PERFORMANCE MODELING AND OPTIMIZATION FOR ON-CHIP INTERCONNECTS IN MEMORY ARRAYS

A Dissertation
Presented to
The Academic Faculty

by

Javaneh Mohseni

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2018

PERFORMANCE MODELING AND OPTIMIZATION FOR ON-CHIP INTERCONNECTS IN MEMORY ARRAYS

Approved by:

Dr. Azad Naeemi, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Oliver Brand
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Muhannad Bakir
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Yogendra Joshi
School of Mechanical Engineering
Georgia Institute of Technology

Dr. Jeffrey Davis
School of Electrical and Computer Engineering
Georgia Institute of Technology

Date Approved: [Month dd, yyyy]

ACKNOWLEDGEMENTS

I wish to thank my family, my great advisers Dr. Azad Naeemi and Dr. James Meindl, my committee members Dr. Jeffrey Davis and Dr. Muhannad Bakir, Dr. Tasha Torrence and Dr. Daniela Staiculescu from the ECE academic office, Chenyun Pan, and the other students of our group.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

In multi-core systems, the memory latency and bandwidth are among the key limitations. While interconnects have created major challenges for the integrated circuit technology in the past decades, there have been major changes in the nature and the severity of the challenges in recent years. Therefore, modeling and benchmarking the interconnect performance for memory chips is of utmost importance. The memory system design is facing many challenges. DRAM-based memory systems are stretched to meet the increasing demands on high memory bandwidth and large memory capacity that are required by multi-core processors. To address these challenges both technology and circuit solutions should be investigated. While this work focuses on a few memory technologies, the modeling approach presented here and the insights obtained regarding the limits and opportunities associated with interconnects apply to other emerging and conventional memory technologies.

CHAPTER 1. INTRODUCTION

The unprecedented exponential growth in the semiconductor industry has been enabled by the dimensional scaling of the silicon-based CMOS technology. Scaling of the transistor leads to lower delay, lower power consumption, higher performance, and higher chip density. This results in integrated circuits (ICs) with higher performance and more functionality. In 1965, Gordon Moore, the co-founder of Fairchild Semiconductor and Intel, made the observation that the number of transistors in a dense integrated circuit doubles every year in his groundbreaking paper "Cramming More Components onto Integrated Circuits" [1], and projected this growth rate would continue for at least another decade. In 1975, on the verge of another decade, he made a revision to his forecast by changing the growth rate to doubling every two years [2]. It is said that David House mentioned that since transistor scaling improves the chip performance by increasing both the number and speed of the transistors, the chip performance would double every 18 months [3].

Scaling down the transistor dimensions has resulted in numerous challenges through the years. One of the challenges was related to the transistor's gate dielectric. Traditionally, silicon-dioxide (SiO_2) was used as the transistor's gate dielectric material for decades. As transistors' size has been reduced, the thickness of the gate dielectric has been reduced also to enable higher gate capacitance and therefore, higher gate control over the channel. As the gate dielectric thickness scales below 2 nm, leakage currents due to tunneling increase drastically [4-5]. The increased leakage currents lead to higher power consumption and reduced device reliability. In order to keep the increasing gate leakage current under control, the gate dielectric scaling was stopped for a few

generations. However, thanks to the introduction of strained-silicon technology which improved mobility, transistor performance improvement was still maintained [6-7]. Eventually, replacing gate SiO_2 with a high-k dielectric material resolved the gate dielectric scaling problem [8].

The use of a high-k dielectric material enabled reducing the equivalent electrical thickness of gate dielectric which resulted in better electrostatic control of the channel by the gate while keeping the physical thickness of gate dielectric constant which prevented the quantum-mechanical tunneling of electrons from the gate to the channel from increasing, and thus, kept the gate leakage currents under control. Replacing SiO_2 with a high-k material brought new complexities to the manufacturing process. Silicon dioxide can be formed by oxidizing the underlying silicon, ensuring a uniform, conformal oxide and high interface quality. As a consequence, development efforts have focused on finding a material with a requisitely high dielectric constant that can be easily integrated into a manufacturing process. Materials which have received considerable attention are hafnium silicate, zirconium silicate, hafnium dioxide and zirconium dioxide, typically deposited using atomic layer deposition [9]. In addition, replacing SiO_2 with a high-k material required replacing the polysilicon gate with a metal gate since the polysilicon gate was not compatible with the high-k dielectric [10]. Another major innovation in the semiconductor industry was extending the transistor in the vertical dimension. At the 22nm technology node, the first non-planar CMOS transistor was introduced [11], and different variations of it have been manufactured under the general name of FinFET. In a FinFET transistor, the channel is extended in the vertical dimension (called fin), and the gate is wrapped around the fin to provide a better electrostatic control from the three

sides. Due to having gates on three sides, the FinFET transistor has alternatively been called the tri-gate transistor. FinFET provides a smaller required supply voltage and reduced short channel effects.

Since the devices and interconnects both contribute to the chip performance, in order to reach higher chip performance, devices and interconnects need to be scaled down simultaneously. Therefore, keeping the chip area constant, as the transistors get smaller and their number gets larger, faster and denser interconnects are required. With scaling down interconnects, many challenges arise with regard to the interconnects performance [12]. Due to size effects, the interconnects' resistivity increases quite significantly, and the total interconnects' capacitance increases due to the high density of the interconnects. The constraints and challenges are different for local and global interconnects. These differences along with their precise definitions in memory chips will be investigated extensively in this work. For now, it is important to know that in memory chips, the global interconnects are long interconnects which carry address and data signals from the edge of the chip to the memory bank, and the local interconnects are relatively shorter interconnects inside the memory bank which carry the signals to the memory cells. As a result, the local interconnects' pitch is limited by the memory technology and memory cell size, while the global interconnects' pitch is only limited by the number of metal levels. As a solution to some of the problems, the number of metal levels has been increased through the years [13]. This multi-level interconnect routing offers several advantages. The local interconnects with finer pitch and higher density are fabricated in the lower levels, and their small pitch values helps preventing the number of metal levels from getting too large. The global interconnects are routed on the higher metal levels, and

the possibility to increase their pitch enables smaller memory latency. Another solution was to replace Aluminum (Al) with copper (Cu). Copper has a lower resistivity [13] which contributes to smaller resistance-capacitance (RC) delay of interconnects. It also has a higher resistance to electromigration [14]. Electromigration is the gradual movement of the atoms in a conductor caused by the momentum transfer from the electrons when a large current is passing through the conductor. The severity of material transport caused by electromigration depends on the electrical current density. Therefore, scaling down cross-sectional dimensions of wires causes electromigration to become a bigger challenge. Another strategy to allow continued scaling of microelectronic devices has been the use of low-k dielectric materials. As the chip transistors and interconnects have scaled down, the insulating dielectrics that separate the conducting elements have become thinner and the cross-talk effect has become more challenging. Replacing silicon dioxide with a low-k dielectric of the same thickness reduces the parasitic capacitance. This reduced capacitance results in lower RC delay and power dissipation. These new materials, processes, and fabrication methods have made interconnects scaling possible for a number of technology generations. It is believed that another solution that will drive the scaling down of the microchips in the next decade is innovations in extending the device integration along the vertical dimension.

The innovative solutions in the semiconductor industry are the results of major investments in research and development. Although extending the device and the chip in the vertical dimension is expected to govern the technological advancement in the near future, the semiconductor industry will have to keep on facing new challenges to continue scaling in the foreseeable future. Among the processing techniques, until the extreme

ultraviolet light (EUV) lithography, which makes use of light at a wavelength of 13nm, becomes available, the industry has to extend the use of 193-nm immersion lithography tools to ultra-scaled technology nodes through optimized multiple-patterning and computational-lithography tools. In this thesis, we focus on the challenge of interconnects which poses significant limitation to the performance of microchips despite the aforementioned innovations.

The semiconductor industry has made major investments in the research for increasingly radical potential solutions to carry technology advancement through dimensional scaling to beyond conventional CMOS. Many companies have dedicated their research pipelines to emerging device and interconnect technologies, such as carbon–based devices [15, 16] and interconnects [17, 18], nano–electromechanical systems (NEMS) [19], optical or photonic interconnects [20, 21], and even non–charge–based systems [22], to extend Moore's Law to beyond–2020 technology generations. However, since the microchip's performance is determined by the performance of both the devices and the interconnects, the improvements of performance, power dissipation and ease of scaling of devices have to be paralleled by similar advances in the interconnects. Therefore, all the research for novel device solutions has to be accompanied by the investigation of solutions to the accompanying interconnect challenges which have become ever more complex from the performance, energy, reliability, and cost aspects.

In terms of challenges related to interconnects, there have been major changes in the nature and the severity of the challenges in recent years. Here, we take a closer look at the interconnect problem and the resulting performance issues that arise from it.

Interconnect delay problem from the resistance-capacitance perspective

As mentioned before, modern electronic chips have a multilevel interconnection network with interconnects with different pitches routed on different levels. Short interconnects that carry signals between transistors that are relatively close to each other, within a certain functional block, are routed at local interconnect levels with fine pitches for high density. As interconnects get longer, they are made wider and thicker to reduce the associated resistance per unit length; hence delay. Therefore, the multilevel interconnection network not only is a solution to the routability problem, but also helps to reduce the interconnect latency [23].

The aggressive scaling and the increasing number of interconnects have created an increasing resistive and capacitive load to the system which leads to higher interconnect RC delay. Whenever the advancement rates of the components of a system are not the same, the component lagging in performance will become the system's performance bottleneck. That is what has happened in the electronic microchips in the recent years. As the intrinsic device performance has been improved with dimensional scaling, the impact of interconnect RC load and latency on the circuit's speed has become more pronounced. Some projections of the ITRS update in 2017 [24] are tabulated in Table 1 to illustrate the severity of the interconnect latency problem. As seen in the table, the interconnect critical length which is the length at which the interconnect delay becomes equal to an NMOS transistor intrinsic delay is quickly decreasing. This means that the interconnects are becoming the latency bottleneck of the microchip.

Table 1 - Interconnect technology parameter projections related to the latency of interconnects extracted from the 2017 update of ITRS [24]. Calculated metrics are indicated with the * sign.

	2015	**2020**	**2025**
M1 half pitch (nm)	21	12	7
Aspect Ratio	1.9	2	2.2
Cu resistivity (μΩ.cm)	6.61	9.74	15.02
Barrier/cladding thickness for Cu M1 wiring (nm)	1.9	1.1	0.6
*Resistance per unit length for M1 wires, r (Ω/μm)	101	434	1750
NMOS intrinsic delay, τ = CV / I (Multi-gate, MG) (ps)	0.32	0.19	0.12
Capacitance per unit length for M1 wires, c (pF/cm)	1.8-2	1.6-1.8	1.5-1.8
*Distributed RC delay of 1mm M1 wire, $\tau_{int} = 0.4\mathrm{rc}L^2$ (ps)	7676	29512	115500
*Length at which $\tau_{int} = \tau$, (μm)	6.5	2.5	1

The interconnect RC delay is proportional to

$$\tau = \rho\varepsilon\frac{L}{HT}$$

This expression shows that the interconnect delay can be reduced by: (1) reducing metal resistivity (ρ) by using new materials, (2) scaling insulator permittivity (ε), (3) reducing the interconnect length (L) using novel architectures, and (4) reverse scaling metal height

(H), and insulator thickness (T). All the proposed solutions to the interconnect problem through time aim at changing the value of one of these parameters. These include switching to Cu/low-k interconnect technology to lower the ρε product, using multiple core architectures to reduce the maximum global interconnect length, and reverse scaling to increase the metal height or the insulator thickness. Another solution to reducing the long global interconnect length is switching to three-dimensional integration.

Another approach to solving the interconnect scaling problem is changing the physical means of signal transmission by switching from metal interconnects to optical interconnects [25, 26, 27]. Each of these novel solutions brings side problems with itself, such as router power dissipation in many-core architectures. Still, it is undeniable that the nature of the global interconnect problem has changed as a result of these advances.

Furthermore, the nature of the interconnect problem has also changed at the local interconnect level, especially for sub-20nm technology nodes. At those ultra-small dimensions, the copper resistivity increases drastically as a result of size effects. The dimensions of within core interconnects are scaled with technology so that the $\frac{L}{HT}$ term in equation (1) is kept almost constant. As a result, the interconnect resistivity becomes the dominating factor in determining the interconnect intrinsic delay. This radical change in Cu interconnect limitations for ultra-scaled future technology nodes has created the motivation for looking at alternatives interconnect materials and technologies that can replace copper at local interconnect levels, where Cu wire dimensions are ultra-small. Carbon–based interconnects have long been considered as a promising alternative for future nanoscale interconnects due to their long mean free path (MFP), high current carrying capability and high thermal conductivity. There has been major technological

progress in fabricating such interconnects and the rising opportunities in terms of energy and performance [26]. However, there are still many major challenges that must be overcome before they can become commercially viable options.

In addition to the interconnect latency problem, interconnect power dissipation is also a major problem. Figure 1 shows the trends in the transistor count, single-thread performance, frequency, power, and number of cores for microprocessors in the last three decades. The advances in the microprocessor frequency and single-thread performance is partly due to making smaller, less power-hungry transistors, increasing the frequency, and partly the result of exploiting instruction level parallelism in pipelined architectures for increased throughput. These advancements were enabled by the rapid increase in transistor density with technology scaling. At each technology node, using smaller and faster transistors enabled building larger cores that provided higher frequency and higher throughput. Eventually, as a result of higher operational frequency and a larger number of smaller, denser transistors and interconnects which means higher capacitive load, the microchip power dissipation became the limiting factor to building larger cores. As a result, the system frequency was slowed down to keep the chip power dissipation under control. Also, since the interconnect latency improvement is not keeping up with the rate of improvement in the transistor latency, the system total delay is being more and more dominated by the interconnects latency. This also contributes to the slowing down of the historical rate of chip frequency increase. Multi-core architectures have been the solution to increasing the overall performance of microprocessors while managing the power dissipation by parallel computation. In the future multi- and many core architectures

design that will enable the pursuit of the Moore's law, power management will continue to be a major issue.

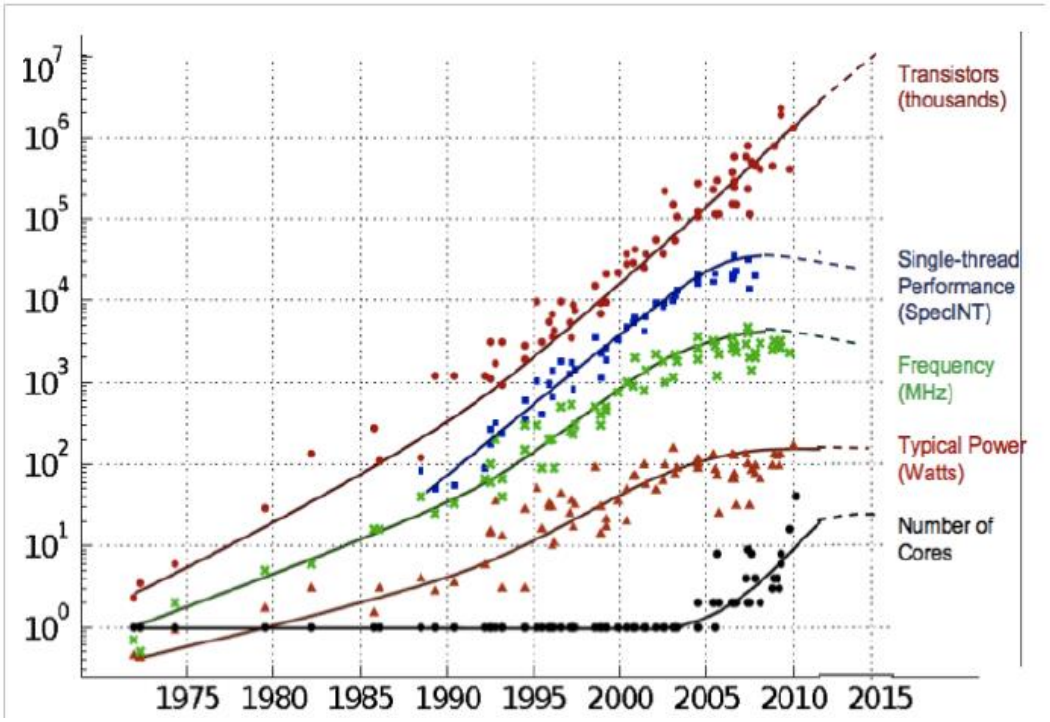


Figure 1 - Microprocessor trend data: The changes in the transistor count, single–thread performance, frequency, power, and number of cores are plotted for the past 35 years. Adapted from [28].

Interconnects account for the source of a large portion of the power dissipated in a microprocessor. An interconnect power analysis study performed on a microprocessor designed for power efficiency, consisting of 77 million transistors, and fabricated in the 0.13 mm technology in 2004, revealed that interconnects account for 50% of the total dynamic power dissipation [29]. Furthermore, with scaling, the interconnect density and hence the total capacitance associated with interconnects increase as well. Lower-k dielectric materials can reduce this capacitance and as a result, the interconnect power dissipation. Table 2 shows the comparison between interconnect and device dynamic

power dissipations at three different technology nodes to underline the significance of the interconnect power dissipation problem.

Table 2 - Interconnect technology parameter projections related to the dynamic power dissipation associated with interconnects extracted from the 2017 update of ITRS [24]. Calculated metrics are indicated with the * sign.

	2015	**2020**	**2025**
M1 half pitch (nm)	21	12	7
Aspect Ratio	1.9	2	2.2
Capacitance per unit length for M1 wires, c (pF/cm)	1.8-2	1.6-1.8	1.5-1.8
NMOS dynamic power indicator per device width, $E = CV^2$ (f J / μ m)	0.42	0.25	0.15
*M1 wire dynamic power indicator per length, $E_{int} = C_{int}V^2$ (f J / μ m)	0.1216	0.079	0.057
*Length at which $E_{int} = E$ for a minimum-width NMOS. (in unit of minimum device width)	3.45	3.16	2.63

Comparisons between interconnect and transistor delay/energy that are shown in Table 1 and Table 2 are performed assuming multi–gate CMOS device and conventional Cu/low–*k* interconnect technology projections. Both device and interconnect parameter projections are industry targets for continued Moore's Law, which may require many innovations to achieve. Therefore, emerging post–CMOS devices that meet these parameter requirements will also suffer from the same limitations imposed by the conventional Cu/low–*k* interconnect. However, most of the current emerging device research is focused on speeding up or reducing the power consumption of a single device.

A simple comparison of the intrinsic gate delay or the dynamic power indicator between a novel device technology and Si–CMOS will not reveal the complete picture of the promise that the new device holds. Interconnect aspects of novel devices have to also be studied for a better understanding of the benefits they may offer.

There have been many publications on the interconnects scaling issues in logic chips [30-31]. However, there has been no comprehensive study on the performance and scalability of interconnects in memory arrays. This thesis presents a comprehensive study of interconnects' performance, reliability, and scalability in different memory technologies, and is divided into 8 chapters.

CHAPTER 2. MEMORY STRUCTURE AND INTERCONNECTS

2.1 Introduction

In this chapter, the structure of the memory system used in the models is broken down into its constituents and explained in detail. Different kinds of interconnects in the memory array are introduced. The chip layout and the metal levels are demonstrated. Different decoding methods are investigated and compared comprehensively. Memory logic circuits are studied, modeled, and parametrical models are presented for the memory chip footprint area.

2.2 Global Interconnects

The memory array is divided into a number of banks. The global interconnects transmit address and data from the array input to the bank, forming a hierarchical-tree (h-tree) network, as shown in Figure 2.

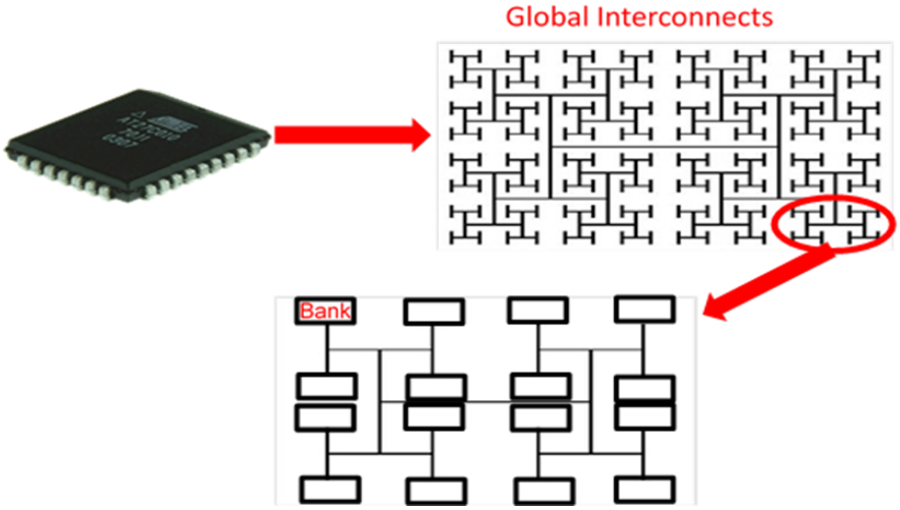


Figure 2 - Memory array structure and global interconnects.

Global Interconnect parameters that affect the memory performance and number of metal levels are wire pitch and length. The memory array is divided into equally-sized segments called memory blocks. Address wires are the interconnects that carry the address of the designated memory block that needs to be read or written to. The data wires are the interconnects that carry the data that is to be written to or has been read from the designated memory cell. Since a memory block is commonly 64 Bytes or 512 bits, the number of data wires is far larger than the address wires. The total length of the global interconnects, which is the length of an interconnect from the edge of the memory array to the edge of the memory bank, is the same for address and data interconnects since they are both transmitted using similar h-tree networks.

2.3 Local Interconnects

The local interconnects are placed inside the memory banks, and are divided into three groups. Each memory bank consists of a number of rows and columns of memory cells. The access to a memory cell is provided via an access transistor. For each row of memory cells, there is a wordline which is connected to the access transistors of all the memory cells in that row. As a result, when a wordline is activated, all the access transistors in the row connected to that wordline are turned on.

For each column of memory cells, there are two interconnects placed along the y-direction that are named bitline (BL) and sourceline (SL). The sourceline is connected to the drain of the cell's access transistor, and the bitline is connected to the cell's value retaining element (e.g., a capacitor in the case of DRAM), on one end, and a sense amplifier on the other end. As shown in Figure 3, this cell value retaining element is placed between the access transistor's source and the bitline.

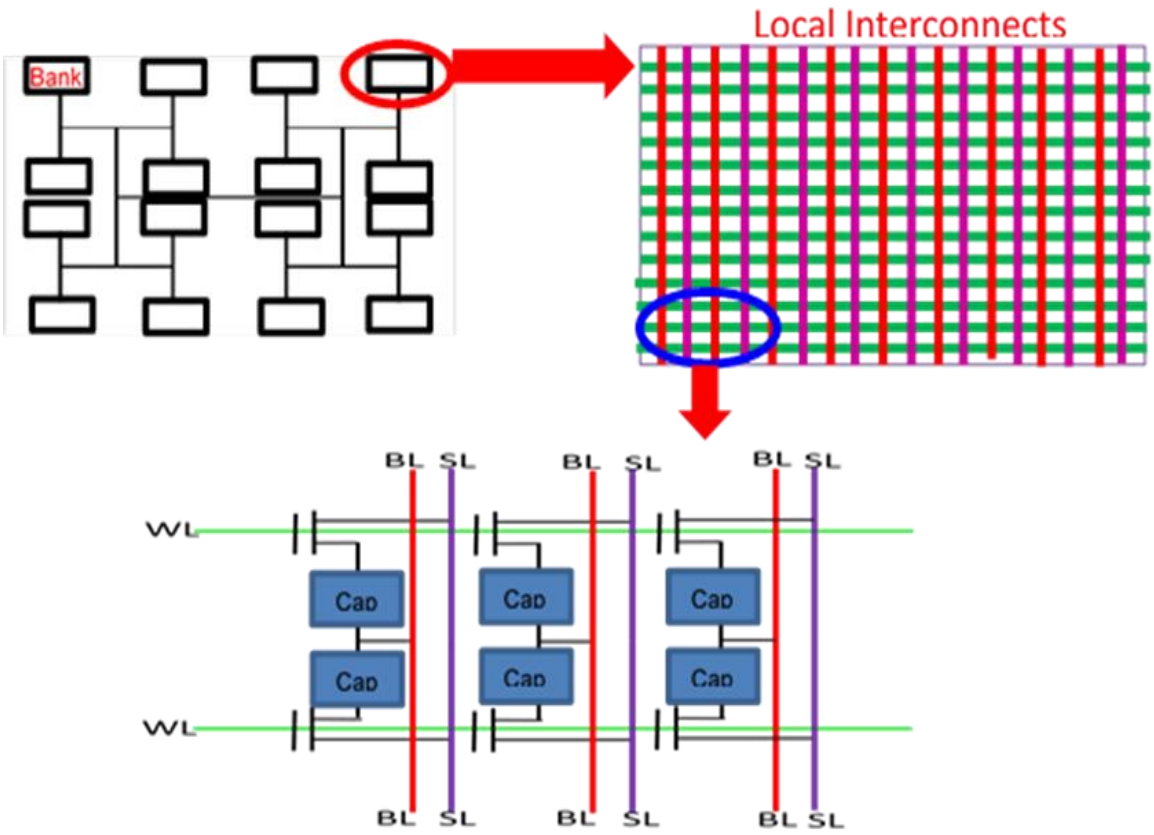


Figure 3 - Memory bank structure and local interconnects.

2.4 Memory Logic Components

The on-chip logic components in a memory array are comprised of decoders, multiplexers, and sense amplifiers. The decoders are built at two levels. The first level decoders choose the designated bank, and the second level decoders choose the designated row inside the bank, as shown in Figure 4. Later in the chapter, memory decoding would be discussed extensively.

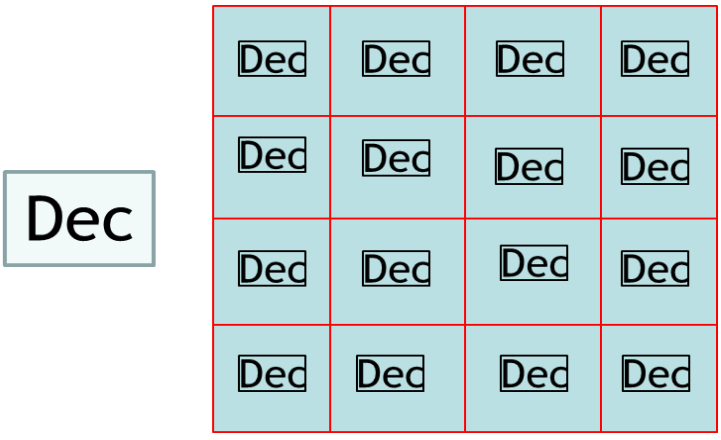


Figure 4 - Double-level memory array address decoders.

There is a sense amplifier connected to every column in every memory bank. There are two levels of multiplexers. As shown in Figure 5, two levels of multiplexers are needed to transmit the data to the memory's output ports. First, there is a multiplexer at each bank to choose the designated columns among all that bank's columns and transmit their data to the edge of the memory. At the next level, the outputs of all the banks' multiplexers are connected to a second multiplexer that chooses and transmits the data coming from the designated bank among all the banks.

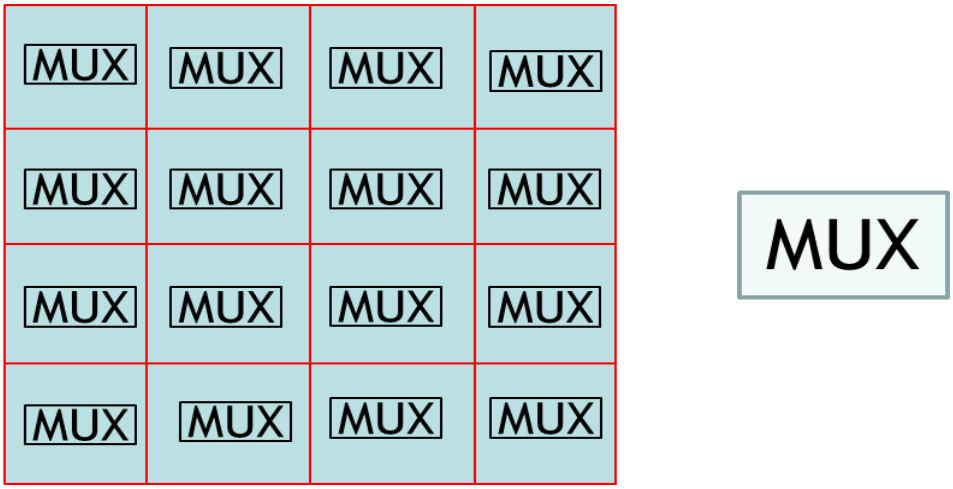


Figure 5 - Double-level memory array data sense amplifiers.

2.5 Multi-Level Interconnect Routing in Memory Arrays

By dividing the interconnects into a number of groups and fabricating them on different metal levels with different wire pitches, the delay and power consumption of the interconnects, and thus the chip area and performance of the memory array, can be optimized. As seen in Figure 6, the first three metal levels are for WLs, BLs, and SLs. In

the lower metal levels with minimum wiring pitch, wires can be routed only in one direction. As a result, WLs cannot be fabricated on the same metal level as the BLs and SLs. BLs and SLs can be fabricated on the same or different metal levels depending on the constraints on the local interconnects pitch and the total number of metal levels. The upper metal levels are for the global interconnects, and their number may vary. Figure 6 shows the cross-sectional view of a memory chip.

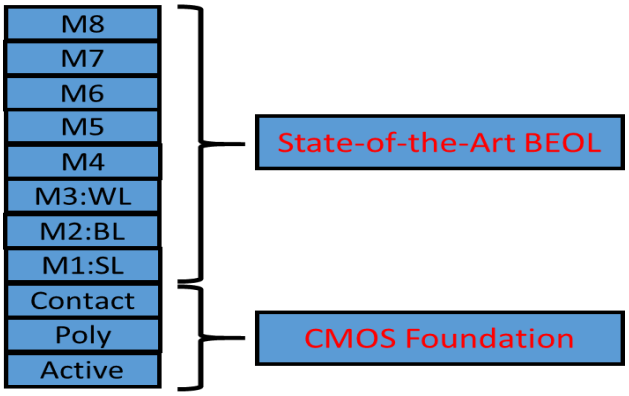


Figure 6 - Memory chip's cross-section view.

Routing local and global interconnects on multiple metal levels enables the optimization of chip performance by routing the fine-pitch local interconnects on the lower metal levels in order to reduce the wiring area and chip dimensions, and placing the wider-pitch global interconnects on multiple upper metal levels in order to increase the cross-sectional dimensions of wires and reduce their delay. On the other hand, adding additional metal levels is difficult due to fabrication challenges. As a result of this tradeoff, the number of a chip's metal levels is an important parameter in memory design. Here, the total number of a memory chip's metal levels is found parametrically, and enables the memory designer to determine the optimum number of metal levels depending on the design specifications and desired criteria. In the calculations, the horizontal side refers to the side along the x-axis, and the vertical side refers to the side

along the y-axis. We make the assumption of square memory banks and square memory array, but we cannot assume that the memory cells are square. Therefore, we define the two parameters a_{horiz} and a_{vert} such that memory cell horizontal length is equal to a_{horiz} F and the cell vertical length is equal to a_{vert} F, where F is the feature size, i.e. 1/4 of M1 pitch. Using these symbols, the memory array area can be written as AF^2 where A is the product of the number of cells in the memory array, a_{horiz}, and a_{vert}, and the total memory die area can be written as $(1+B)AF^2$ where B is the percentage added area due to logic circuits. As a result,

$$L_{die} = \sqrt{(1+B)A}\,F$$

The next step in finding the number of metal levels is finding the required total wiring area. In order to simplify the math, it is assumed that individual (final) decoding is used for all the parts of a memory cell's address; i.e. address decoding is done individually at each bank. Different possible methods of address decoding are discussed later in the chapter. In order to be able to use the h-tree network for the global interconnects, the number of banks at each die side should be a power of 2.

$$N_{bank} = 2^{2m}$$

where m is a chosen integer number. As a result

$$m = 0.5\,log_2(N_{bank})$$

If the number of banks along the two horizontal and vertical die sides are not assumed to be equal, the number of banks would be 2^{m_horiz} along the horizontal side, 2^{m_vert} along the vertical side, and $2^{m\_horiz+m\_vert}$ in total.

Total length of the h-tree interconnects is found by finding the sum of a geometric series. In order to save space, the final result is written here.

Total length of h-tree interconnects is

$$L_{die}\ (1.5 - 0.5^{m}) = \sqrt{(1+B)A}\ F\ (1.5 - 0.5^{0.5\ log_2(N_{bank})})$$

The number of bits in the global bank address is

$$N_{bank\_address\_wires} = log_2(N_{bank})$$

The number of bits in the global row address is

$$N_{row\_address\_wires} = log_2(N_{bank\_rows})$$

where N_{bank_rows} is the number of rows in a bank. Similarly, the number of bits in the global column address is

$$N_{column\_address\_wires} = log_2(N_{bank\_columns})$$

where $N_{bank\_columns}$ is the number of columns in a memory bank. The total number of bits for the global address is the sum of the three above expressions.

$$N_{global\_address} = log_2(N_{bank}\ N_{bank\_rows}\ N_{bank\_columns})$$

Number of global data wires is

$$N_{global\_data} = S_{block}$$

where S_{block} is the memory block size. The total number of global interconnects is the sum of the numbers of global address and data interconnects.

$$N_{global\_wires} = log_2(N_{bank}\ N_{bank\_rows}\ N_{bank\_columns}) + S_{block}$$

We assume the pitch of global address and data interconnects to be the same. In order to find the number of global interconnects metal levels, we use the h-tree network

to find the global interconnects area routed in the area among the banks. To save space, we only write the final result here. The global interconnects area is found as

$$A_{global\_interconnects} = (N_{bank} - 1)^2 N_{global-wires}^2 P_{global-wires}^2 + 2\sqrt{(1+B)A}\,(N_{bank} - 1) F\, N_{global-wires}\, P_{global-wires}$$

Now, we have to find the number of required global metal levels. We assume that wires can cover half of the die area. As a result

$$N_{global\_interconnects\_metal\_levels} = \frac{[2\,(\sqrt{N_{bank}} - 1)^2\, N_{global-wires}^2\, P_{global-wires}^2 + 4\,\sqrt{(1+B)\,A}\,(\sqrt{N_{bank}} - 1)\, F\, N_{global-wires}\, P_{global-wires}]}{[(1+\mathrm{b})\,\mathrm{A}\,\mathrm{F}^2]}$$

The above equation gives the number of global metal levels as a function of technology node, global wire pitch, memory cell size, total memory capacity, number of memory banks, and memory block size.

Additionally, a memory chip has 3 local interconnects metal levels for WL, BL, and SL.

2.6 Memory Address Decoding

In this section, we describe different stages of memory access, and the set of required interconnects and peripheral circuits that are required for each part of memory operation. There are multiple ways of creating the logic circuits and routing the interconnects required for implementing each part of memory operation, which are

advantageous to one another in terms of wiring area, logic circuits area, memory delay, power consumption, etc.

The first stage of memory access is reading and decoding the address in order to find the designated memory block in the memory array. The memory address is divided into 3 parts which are the bank address, row address, and column address.

2.6.1 Memory Bank Address Decoding

There are three ways to decode the memory bank address and activate the designated memory bank. We call these methods “Central Decoding”, “Individual Decoding”, and “Distributed Decoding”. They could alternatively be named “Initial Decoding”, “Final Decoding”, and “Intermediate Decoding”, respectively.

The first decoding method is central decoding or initial decoding. In this method, the memory bank address is decoded at the edge of memory die, and decoded information is transmitted to the memory banks using an h-tree network. One wire goes to each bank which carries a 1 for the designated bank, and 0 for the other banks. Figure 7 shows the initial bank address decoding for a memory array with 2^2 banks. The small number of banks has been chosen for clarity.

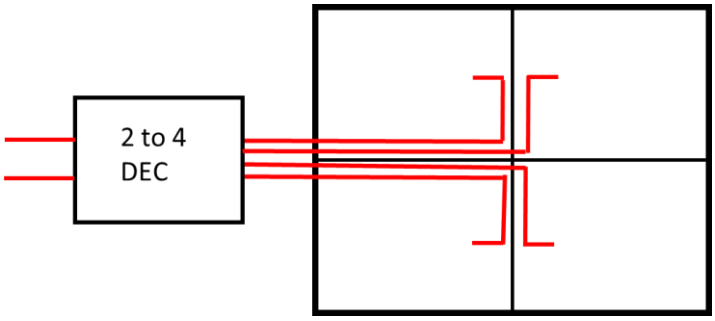


Figure 7 - Initial bank address decoding for a memory array with 22 banks.

The advantages of central decoding for memory bank address are:

1. Smaller logic circuits area: There is one central set of decoders for the bank address instead of a set of decoders at each bank.

2. Lower average interconnects power consumption: This might sound counterintuitive. In this method, the number of wires that go to the banks is larger, but one and only one wire carries a high signal since information is being transmitted in a decoded manner, whereas when bank address is transmitted in a coded manner, although the total number of wires is smaller, more than one wire could be carrying a high signal. For example, coded and decoded information transmission for activating bank number 15 and bank number 1 in two consecutive bank accesses is shown in Figure 8.

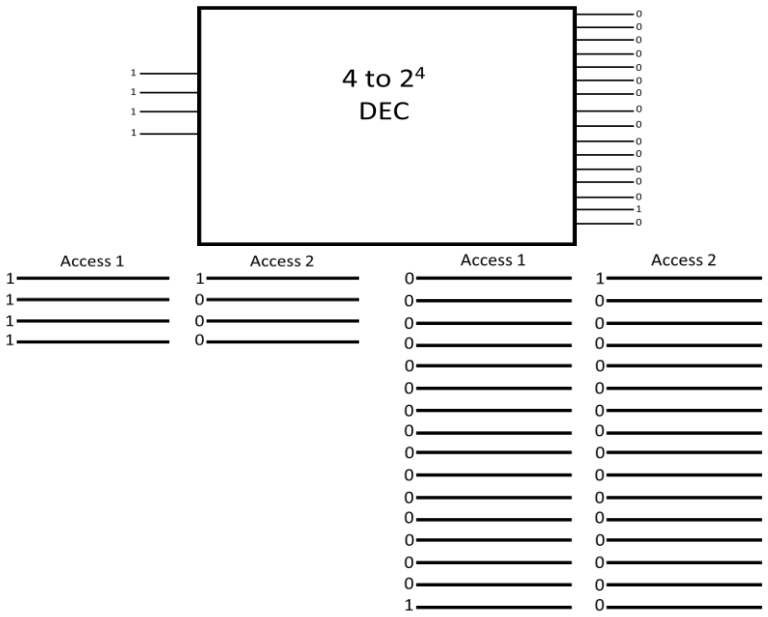


Figure 8 - Coded and decoded information transmission for activating bank number 15 in a memory array with 2^4 banks.

The disadvantage of central decoding for memory bank address is the larger wiring area.

The second decoding method is individual decoding or final decoding. This method is to send the bank address to all the memory banks in a coded manner. For example, for a memory array with 2^{18} banks, 18 wires are routed to all the banks using an h-tree network. There is a partial decoder, built using a number of inverters and an AND gate, at each bank. The decoder output wire is used as the enable signal for the row decoders in that bank. Figure 9 shows the individual (final) bank address decoding scheme for a memory array with 2^2 banks.

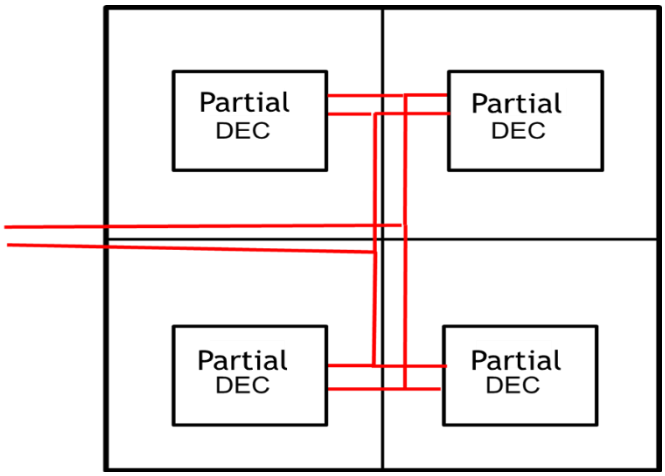


Figure 9 - Individual (final) bank address decoding for a memory array with 2^2 banks.

Advantages:

1. Smaller wiring area. Transmitting the bank address as coded signal to the banks requires a far smaller number of wires.

Disadvantages:

1. Larger logic circuits area. Since only one of the outputs of the full decoder is required as the enable signal at each bank, there is only a partial decoder at each bank. Still, since

in addition to one AND gate, there is a number of inverters at each bank, the total logic circuits area is larger than the case with one central decoder.

2. Higher interconnects power consumption. This is opposite of the case for lower interconnects power consumption using central decoding.

The third decoding method is distributed decoding or intermediate decoding. In method 1, the bank address is decoded at the beginning of the path using a central decoder which is common among all the banks. The disadvantage of this method is very significant since transmitting address in a decoded manner to the banks could increase the number of required metal levels for a memory chip to up to 30 levels, which is far more than the common value for fabricated memory chips. In method 2, address is transmitted to the banks in a coded manner, and is decoded using individual decoders at each bank. This method increases the required decoders area significantly, and is an important issue since among the logic circuits components in memory array, which mainly consist of decoders, sense amplifiers, and multiplexers, decoders are the main contributors to the total logic circuits area in memory arrays. The common practice is that the total logic circuits area should not exceed 10% of the total memory cells area, and keeping the area of these logic circuits below this limit is already a challenge.

In response to the challenges resulting from central and individual decoding techniques, a third option, named distributed decoding, is proposed for address decoding in memory arrays. In this method, one step of decoding is done at each intersection along the wire path in the h-tree network. As a result, the information is completely decoded by the time it reaches each memory bank, and only a single 0 or 1 signal is delivered to each

memory bank. This signal is used as the enable signal for the row decoder inside that memory bank.

Figure 10 shows the distributed decoding method for a memory array with 2^6 banks by showing only the signal carrying global interconnects. Figure 11 shows this decoding method by showing all the global interconnect wires. The comparison between these two figures shows the small ratio of signal carrying interconnects using this decoding method which results in major power dissipation reduction in global interconnects.

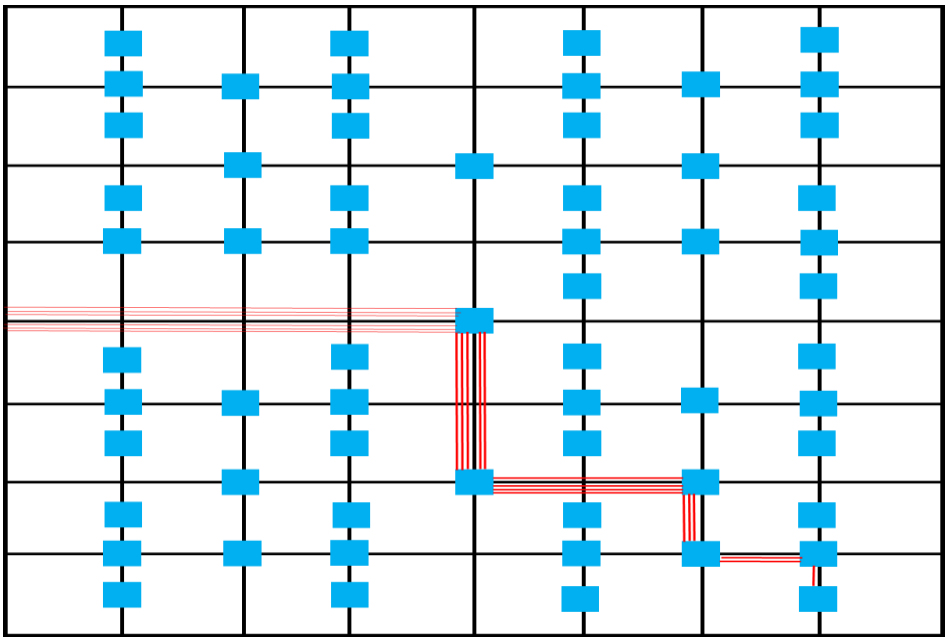


Figure 10 - Distributed (Intermediate) decoding method for bank address decoding for activating bank number 63 in a memory array with 2^6 banks by showing only the signal carrying global interconnects.

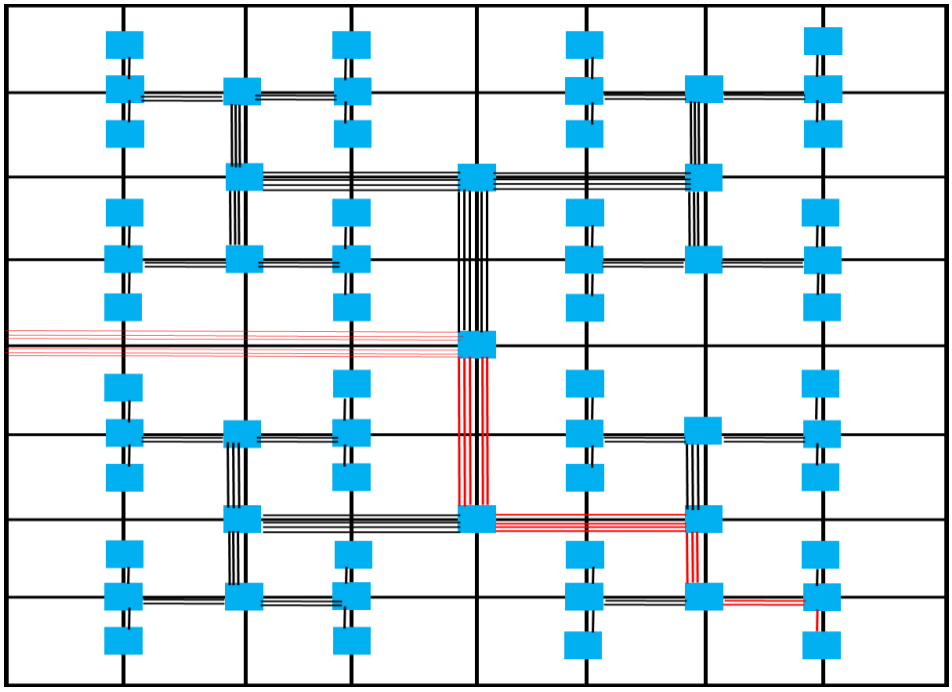


Figure 11 - Distributed (Intermediate) decoding method for bank address decoding for activating bank number 63 in a memory array with 2^6 banks by showing all the global interconnect wires.

As stated before, the three goals in the interconnects design in memory arrays are reducing the wiring area, reducing the logic circuits area, and improving the memory performance by reducing the delay and/or dynamic power consumption. Usually, in the memory design, there is a trade-off between these desired results. However, for the case of decoding the bank address, distributed decoding satisfies all of these goals simultaneously. The reasons that this is possible are:

1. Magic number 2. Since the bank address in a memory array is transmitted as a binary number when transmitted as coded signal, the most significant bit (MSB) corresponds to one half of the memory cells. The second MSB corresponds to the two halves in each half of the array, and so on. On the other hand, in the h-tree network, at the

first intersection each signal is copied into two signals that go to the two array halves at each side. This similarity of dividing into 2 at each step along the wire path creates great convenience for decoding the bank address. At each h-tree intersection, a logic circuit reads the bank address MSB, and, depending on the value of the MSB, directs the bank address to the desired half of the memory array.

2. Most of the time, decoding an address results in an increase in the number of required wires to carry the decoded signal. However, in the case of decoding the bank address, as explained in the previous point, the result of decoding the bank address is directing the signal to the designated memory half array. Not only the number of wires does not increase, but it is also reduced by 1 at each h-tree junction.

3. At each h-tree section instead of the signal being copied into two signals which go to the two array halves at each side, the signal only carries on to the half array which contains the designated memory bank. This way, by limiting the signal to only half of the array at each junction, considerable dynamic power is saved.

2.6.2 Memory Row Address Decoding

For transmitting the row address inside the memory bank, there are essentially only two methods which are central decoding, and individual decoding.

First, the central decoding for memory row address is discussed. Similar to the case of the bank address, in the central decoding method, the row address is decoded at the edge of the memory array, and transmitted to the banks as decoded information using a large number of wires, which take up a large percentage of the die area.

The other option is using individual decoding for memory row address. In the individual decoding technique, the row address is transmitted to the banks in the coded format, and decoded at the bank. This results in large logic circuits area since each bank has its own set of decoders for the row address.

Finally, we talk about a combination of individual and distributed decoding for memory row address. The distributed decoding is quite different for the row address from the bank address. In the case of the bank address transmission, the information that is decoded and used for choosing the signal path at each h-tree junction is the same information that is being transmitted. However, for the row address transmission, the signal used for choosing the signal path at each h-tree junction is the bank address, while the transmitted signal is the row address inside the bank. As a result of this, decoding the transmitted signal (row address) along the signal path results in the increase of wires for carrying the decoded signal. This is yet another case of the trade-off between wiring area and logic circuits area in the memory interconnects design. Since, as mentioned before, the issue of large wiring area is more critical than the logic circuits area in memory chips, the individual decoding method is chosen for the row address. However, in order to direct the row address signal along its path to the designated memory bank, distributed decoding for the bank address could be added to the individual decoding method for the row address. At each row address interconnects h-tree junction, the output of the logic circuits that directed the bank address signal are also used to direct the row address to the half array which contains the designated bank. As explained before, this results in significant dynamic power dissipation reduction in the global interconnects.

2.6.3 Memory Column Address Decoding

Column address decoding could be done with the same methods as row address decoding. There are central (initial) decoding where the bank column address is decoded at the edge of the memory array, and decoded column address is transmitted to the banks. This method has the disadvantage of large required wiring area. The second method is individual (final) decoding, where the column address is transmitted to the banks in a coded format, and is decoded by individual decoders at each bank. This method has the disadvantage of large decoders area since each bank has its own set of decoders. Despite the disadvantage of the second method, since distributed decoding is not possible for the row and column address, individual (final) decoding is used for the row and column address decoding. However, as mentioned before, the method of individual decoding for row and bank address is combined with distributed decoding, where bank address distributed decoding is used to direct the coded row and column address signals along their path in the h-tree network.

In conclusion, the optimized address decoding in memory arrays is done as below:

Memory address part I-bank address → Distributed decoding

Memory address part II-row address → Individual and distributed decoding combination

Memory address part III-column address → Individual and distributed decoding combination

This optimized address decoding structure results in minimized wiring and logic circuits area, and improved performance.

2.7 Memory Logic Circuits with Optimized Decoding Structure

First, we find the total area of decoders for a memory array using any memory technology.

In the distributed decoding method for memory bank address, there is only one decoder which has been divided into several stages and distributed along the signal path. As a result, in this method the required decoder for bank address is one $[log_2(N_{bank})\ to\ N_{bank}]$ decoder. In memory arrays, decoders are built using several [2to4] decoders and NAND gates as shown in Figure 12.

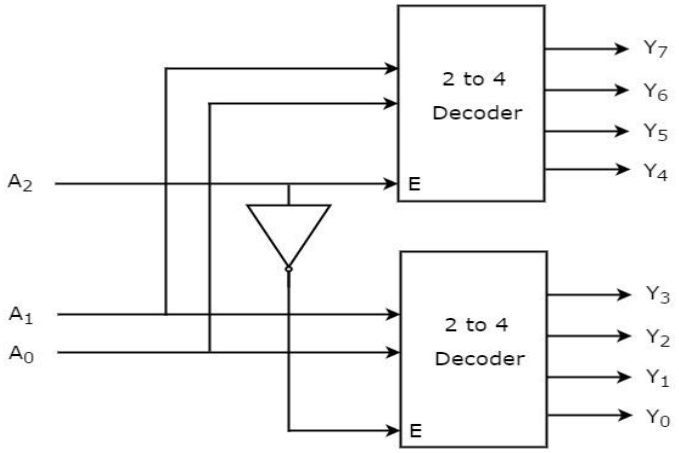


Figure 12 - A 3 to 8 decoder built using three levels of 2 to 4 decoder building blocks.

The area of a [Ato2^A] decoder using [2 to 4] decoder and NAND gate building blocks can be found as

$$A_{[Ato2^A]\,DEC} = \left\lfloor \frac{4^{\log_4 2^A} - 1}{3} \right\rfloor \times A_{[2to4]\,DEC}$$

So, the area of one $[log_2(N_{bank})\ to\ N_{bank}]$ decoder for decoding the memory bank address is found as

$$A_{bank\_address\_DEC} = \left\lfloor \frac{4^{\log_4 N_{bank}} - 1}{3} \right\rfloor \times A_{[2to4]\ DEC}$$

The area of 2to4 decoder is about $1300F^2$, therefore,

$$A_{bank\_address\_DEC} = 1300F^2 \left\lfloor \frac{4^{\log_4 N_{bank}} - 1}{3} \right\rfloor$$

For the memory row address, we use individual decoding, and we use the result of distributed decoding for bank address to direct the coded row address signal along its path in the h-tree network. Therefore, there is a $\left[log_2\left(N_{bank\_rows}\right)\ to\ N_{bank\_rows}\right]$ decoder in each bank.

$$N_{bank\_rows} = \frac{L_{array}}{a_{vert}\ F\ \sqrt{N_{bank}}}$$

where $a_{vert}\ F$ is the memory cell vertical side length. As a result,

$$A_{row\_address\_DEC} = \left\lfloor \frac{4^{\log_4 \frac{L_{array}}{a_{vert}\ F\ \sqrt{N_{bank}}}} - 1}{3} \right\rfloor \times A_{[1to2]\ DEC}$$

$$A_{row\_address\_DEC} = N_{bank} \left\lfloor \frac{4^{\log_4 \frac{L_{array}}{a_{vert}\ F\ \sqrt{N_{bank}}}} - 1}{3} \right\rfloor \times A_{[1to2]\ DEC}$$

$$= 1300\ F^2\ N_{bank} \left\lfloor \frac{4^{\log_4 \frac{L_{array}}{a_{vert}\ F\ \sqrt{N_{bank}}}} - 1}{3} \right\rfloor$$

For the memory column address, we use individual decoding for column address, and we use the result of distributed decoding for bank address to direct the coded column address signal along its path in the h-tree network. Therefore, there is a $\left[log_2\left(N_{bank\_columns}\right)\ to\ N_{bank\_columns}\right]$ decoder in each bank.

$$N_{bank\_columns} = \frac{L_{array}}{a_{horiz}\ F\ \sqrt{N_{bank}}}$$

where $a_{horiz}\ F$ is the memory cell horizontal side length. As a result,

$$A_{column\_address\_DEC} = \left\lfloor \frac{4^{\log_4 \frac{L_{array}}{a_{horiz}\ F\ \sqrt{N_{bank}}}} - 1}{3} \right\rfloor \times A_{[1to2]\ DEC}$$

$$A_{column\_address\_DEC} = N_{bank} \left\lfloor \frac{4^{\log_4 \frac{L_{array}}{a_{horiz}\ F\ \sqrt{N_{bank}}}} - 1}{3} \right\rfloor \times A_{[1to2]\ DEC}$$

$$= 1300\ F^2\ N_{bank} \left\lfloor \frac{4^{\log_4 \frac{L_{array}}{a_{horiz}\ F\ \sqrt{N_{bank}}}} - 1}{3} \right\rfloor$$

The memory address was divided into 3 parts, optimized decoding method for each part was presented, and the area of decoders were found. The total decoder area in a memory array is found as

$$A_{DEC} = A_{bank\_address\_DEC} + A_{row\_address\_DEC} + A_{column\_address\_DEC}$$

$$= \left[\left[\frac{4^{log_4^{N_{bank}}} - 1}{3} \right] + N_{bank} \left[\frac{4^{log_4^{\frac{L_{array}}{a_{vert} F \sqrt{N_{bank}}}}} - 1}{3} \right] \right.$$

$$\left. + N_{bank} \left[\frac{4^{log_4^{\frac{L_{array}}{a_{horiz} F \sqrt{N_{bank}}}}} - 1}{3} \right] \right] \times 1300F^2$$

Next, memory sense amplifiers are discussed. Each column in a memory bank is connected to a sense amplifier; hence, to find the total number of sense amplifiers, we just need to find out the total number of columns in all the memory banks of an array. The number of columns in each bank was found in the previous section. By multiplying $N_{bank\_columns}$ by the number of banks, the total number of sense amplifiers is found.

$$N_{sense\_amplifiers} = \frac{N_{bank}\, L_{array}}{a_{horiz}\, F}$$

The size of each sense amplifier is about $600F^2$, so the total size of sense amplifiers in an array is

$$A_{sense\_amplifiers} = 600F^2\ \frac{N_{bank}\, L_{array}}{a_{horiz}\, F}$$

Finally, we talk about the multiplexers in a memory array. We define the R parameter as

$$R = \frac{N_{bank\_columns}}{S_{block}}$$

To transmit the signal for each bit of the desired memory block, one signal should be chosen among a number of signals equal to R. In memory arrays, multiplexers are built using several stages of small 2to1 multiplexers as shown in Figure 13.

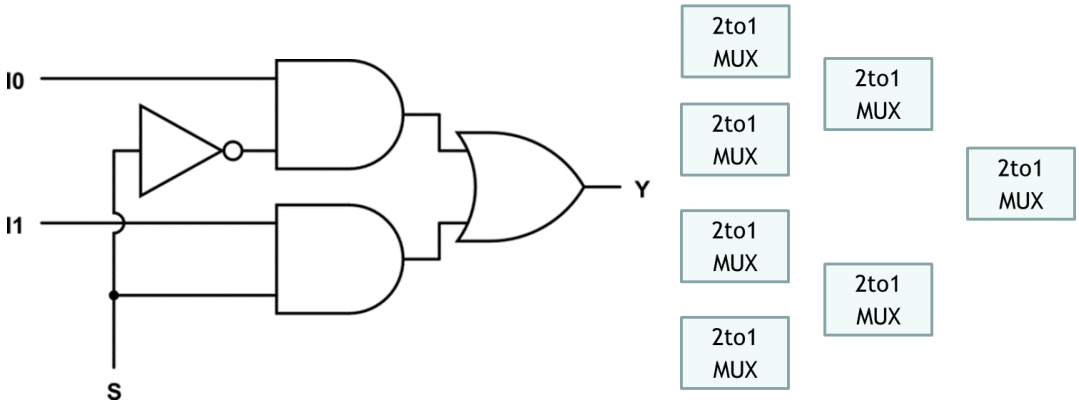


Figure 13 - A 2 to 1 multiplexer circuit and an 8 to 1 multiplexer built using three levels of 2 to 1 multiplexers.

The area of a [Ato1] multiplexer using 2to1 multiplexer building blocks can be found as

$$A_{[Ato1]\,MUX} = \lfloor 2^{lnA} - 1 \rfloor \times A_{[2to1]\,MUX}$$

There are S_{block} number of [R to 1] multiplexers in each bank. As a result, the total area of memory array multiplexers is

$$A_{MUX} = N_{bank}\, S_{block} \left\lfloor 2^{\ln R} - 1 \right\rfloor \times A_{[2to1]\,MUX}$$

The area of 2to1 multiplexer is about $900F^2$, so

$$A_{MUX} = 900F^2\, N_{bank}\, S_{block} \left\lfloor 2^{\ln R} - 1 \right\rfloor$$

We found the area of decoders, sense amplifiers, and multiplexers. The total logic circuits' area in a memory array would be the sum of them.

$$A_{logic} = A_{DEC} + A_{sense\_amplifiers} + A_{MUX}$$

$$= \left[\left[\frac{4^{log_4^{N_{bank}}} - 1}{3} \right] + N_{bank} \left[\frac{4^{log_4^{\frac{L_{array}}{a_{vert} F \sqrt{N_{bank}}}}} - 1}{3} \right] \right.$$

$$\left. + N_{bank} \left[\frac{4^{log_4^{\frac{L_{array}}{a_{horiz} F \sqrt{N_{bank}}}}} - 1}{3} \right] \right] \times 1300F^2 + 600F^2 \frac{N_{bank} L_{array}}{a_{horiz} F}$$

$$+ 900F^2 N_{bank} S_{block} \left[2^{\ln R} - 1 \right]$$

2.8 Conclusion

In this chapter, the structure of the memory system used in the models was broken down into its constituents and explained in detail. Different kinds of interconnects in the memory array were introduced. The chip layout and all the metal levels were demonstrated. Different decoding methods were investigated and compared comprehensively. Memory logic circuits were studied, modeled, and their footprint area was found parametrically. In the next chapter, the DRAM array performance is investigated using the model developed in this chapter.

CHAPTER 3 DRAM MEMORY ARRAYS

3.1 Introduction

Dynamic random-access memory (DRAM) is a type of random-access memory that stores each bit of data in a separate capacitor within an integrated circuit. The capacitor can either be charged or discharged; these two states are taken to represent the two values of a bit, conventionally called 0 and 1. Since even "nonconducting" transistors always leak a small amount, the capacitors will slowly discharge, and the information eventually fades unless the capacitor charge is refreshed periodically. Because of this refresh requirement, it is a dynamic memory as opposed to static random-access memory (SRAM) and other static types of memory. Unlike flash memory, DRAM is volatile memory (vs. non-volatile memory), since it loses its data quickly when power is removed. However, DRAM does exhibit limited data remanence.

DRAM is widely used in digital electronics where low-cost and high-capacity memory is required. One of the largest applications for DRAM is the main memory (colloquially called the "RAM") in modern computers; and as the main memories of components used in these computers such as graphics cards (where the "main memory" is called the graphics memory). In contrast, SRAM, which is faster and more expensive than DRAM, is typically used where speed is of greater concern than cost, such as the cache memories in processors.

The advantage of DRAM is its structural simplicity: only one transistor and a capacitor are required per bit, compared to four or six transistors in SRAM. This allows DRAM to reach very high densities. The transistors and capacitors used are extremely

small; billions can fit on a single memory chip. Due to the dynamic nature of its memory cells, DRAM consumes relatively large amounts of power, with different ways for managing the power consumption.

3.2 Modeling Approaches and Assumptions

We start by studying a DRAM chip at the technology generation of 9.5nm as defined by the ITRS 2013. The area of the chip is assumed to be $100mm^2$ which is equivalent to 7.3GB memory capacity assuming a cell size of $8F^2$ [24]. The size of the memory block is 64B. The memory array is divided into segments called banks. There is a certain percentage of added area due to the peripheral circuits and its value is determined by the number of banks. At the beginning, all the wires are assumed to be minimum size. Different schemes of wiring are investigated later. Both address and data are transmitted through hierarchical tree (H-tree) networks. Each bank has its own decoders, sense amplifiers and multiplexers. Decoders and multiplexers are built using consecutive stages of 2 to 4 decoders and 2 to 1 multiplexers, respectively. Elmore delay model is used to calculate the interconnect delay. An optimal number of repeaters are used to lower the delay of the global interconnects. For the wordlines and bitlines, no repeaters are used due to the cell size limit. To calculate the delay of the wordlines, the capacitance of transistors connected to the wordlines is added to the wire capacitance. To find the cell refreshment power, the power used to read the cell value and recharge the cell storage capacitor is calculated as well.

3.3 Model Results and Discussions

In this section we show the results of the models for access time, dynamic power, access energy, and energy-delay product (EDP). Since the main focus of this work is on

the impact of on-chip interconnects, the dynamic power modeled and presented in this work is only for the individual memory array, which does not include the peripheral I/O circuitry, pipeline clocking, control logic, and off-chip interconnects.

3.3.1 Access Time, Dynamic Power, EDP, and Area Models

As seen in Figure 14, the interconnect delay dominates the overall delay. At a small number of banks, since the bank size is large and the wordline is long, most of the delay comes from the wordline. As the number of banks increases and banks get smaller, the main source of the memory delay becomes the delay associated with the address and data interconnects. As seen in Figure 15, data interconnects account for most of the dynamic power because 512 data interconnects are connected to each bank.

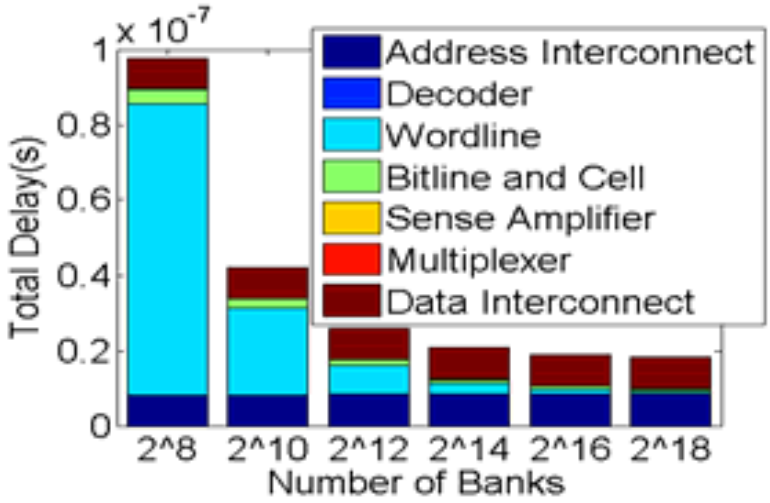


Figure 14 - Memory access time components.

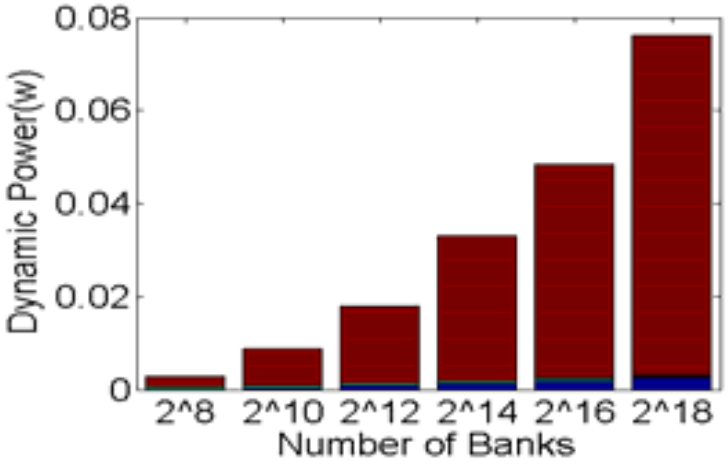


Figure 15 - Memory dynamic power consumption components.

Figure 16 shows the memory access energy and EDP. There is an optimal number of banks that minimizes the EDP. Figure 17 shows the additional area due to the peripheral circuits such as decoders, sense amplifiers, and multiplexers. As the number of banks increases, the total number of each of these components increases because each bank has its own dedicated peripheral circuits, causing an increase in the total area of peripheral circuit components, shown in Figure 17.

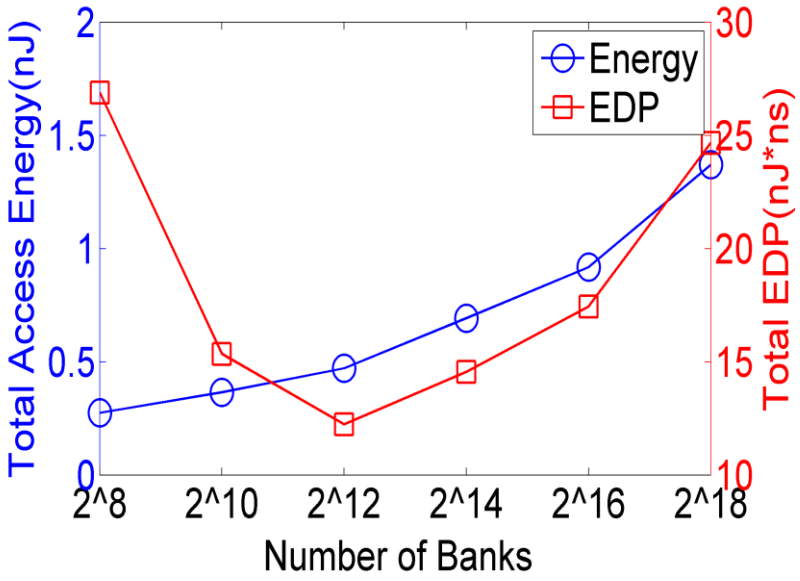


Figure 16 - Memory access energy and EDP.

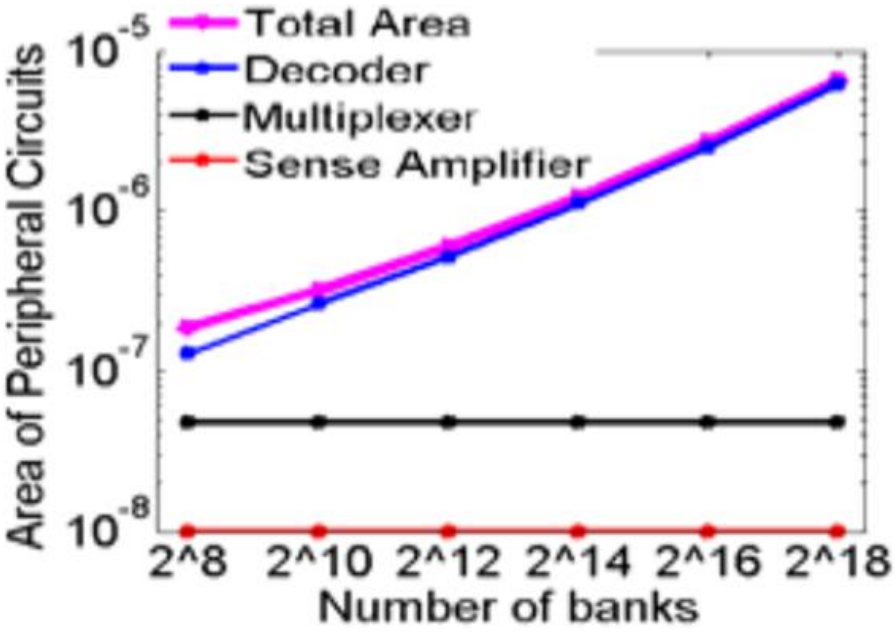


Figure 17 - Memory peripheral circuits' area.

3.4 Interconnects Optimization

In this section we study different wiring and memory organization schemes aimed at lowering the interconnect delay, power, and area.

3.4.1 Adding Interconnect Levels

We compare the memory access times for cases where address and data interconnects are routed in two, three, and four metal levels. As seen in Figure 18, with an increase in the wire pitch and the number of metal levels, the delays of the address and data interconnects decrease. The delays of the wordline and bitline; however, remain unchanged because the wiring pitch for those wires is fixed due to the cell size limit at the 9.5nm technology node.

One of the major components of the memory access time is the wordline delay. This is because of the large resistance and capacitance associated with wordline wires

and the large capacitive load that wordlines have to drive (i. e. the transistors connected to the wordline). Conventional delay mitigation techniques such as repeater insertion and reverse scaling cannot be applied to the wordline because of the limit imposed by memory cell size.

3.4.2 Increase of Decoder Drive Current

As seen in Figure 19, by increasing the size of the decoder output drivers to twice the minimum size and doubling their drive current, the wordline delay is reduced by 21%. After that, the reduction is not notable as the wire resistance becomes dominant. This indicates that the role of wire resistance is much bolder in its contribution to the wordline delay than the driver resistance.

3.4.3 Optimize Bank Aspect Ratio

In the figures, the wordlines inside the banks are always shown along the x-axis, and the bitlines are shown along the y-axis. The bank's aspect ratio is defined as the ratio of its length along the y-axis to its length along the x-axis, i.e. the ratio of the bitline length to the worline length. As seen in Figure 20, by increasing the bank aspect ratio the wordline length and delay decrease but the bitline length and delay increase. As a result, an optimal aspect ratio exists which is equal to 2:1.

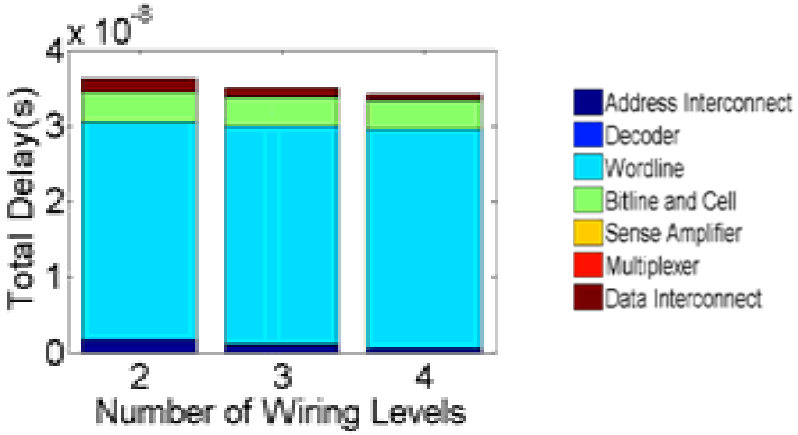


Figure 18 - Impact of various number of wiring levels for data interconnects.

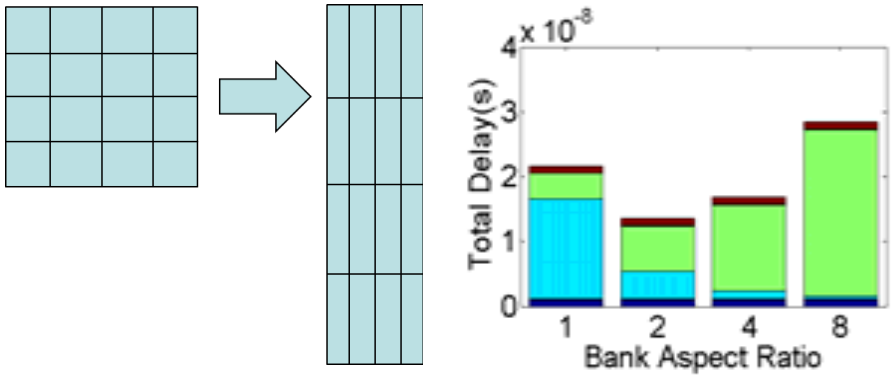


Figure 19 - Impact of banks aspect ratio on memory access time.

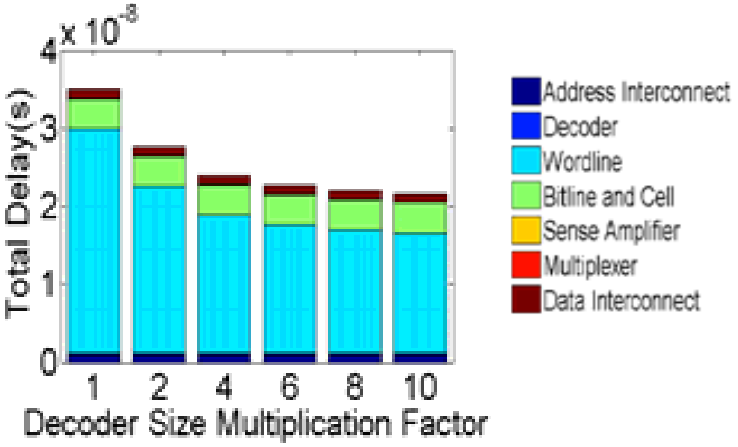


Figure 20 - Impact of decoder size on memory access time.

3.5 Interconnect Technology Solutions

3.5.1 Single Crystal Copper Interconnect

A range of potential technological solutions are being pursued to address the interconnect problem. One promising approach is to switch to subtractive patterning methods to increase the grain size in copper interconnects and eventually to approach single crystal copper interconnects [32]. We have studied the impact of this potential grain size growth on the interconnect delay and total access time. Figure 21 shows the results for the case of the 7nm technology node, where up to 32% improvement is gained in speeding up the memory by fabricating single crystal copper interconnects. It can also be seen that once the grain size becomes larger than 6 or 8 times the wire width there is a diminishing return in delay reduction as surface scattering becomes dominant.

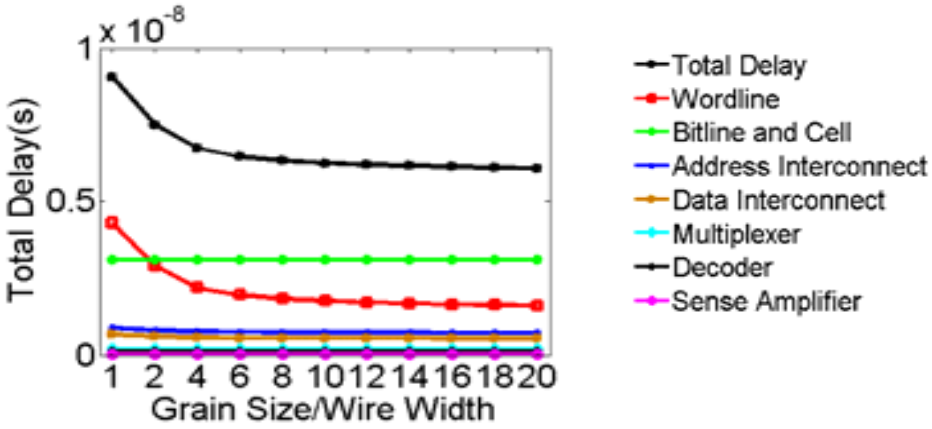


Figure 21 - Impact of grain size of copper interconnect on memory access time.

3.5.2 Changing Barrier Material Fabrication

Another potential improvement in the copper interconnect technology is reducing the barrier thickness and ultimately approaching a barrier-less copper interconnect technology. We have studied the impact of various barrier thicknesses for the memory chip at the 7nm technology node, shown in Figure 22. Barrier thickness reduction has a bolder effect at technology nodes below 10nm.

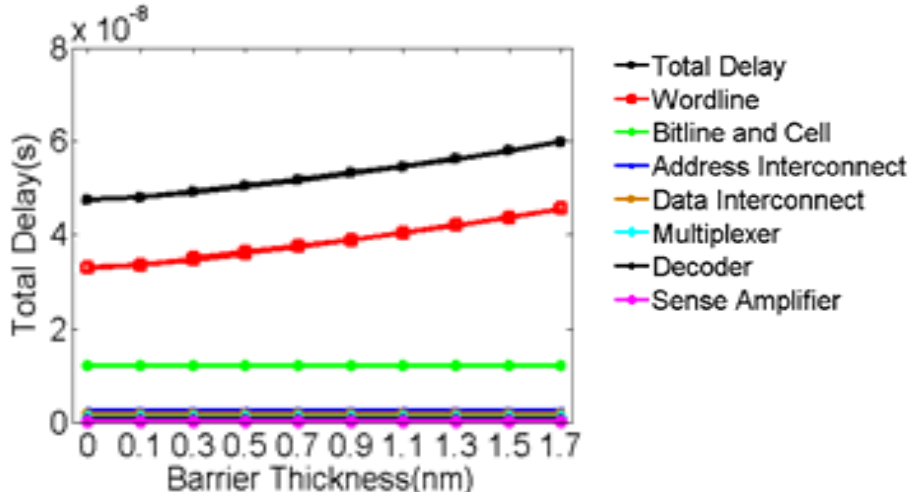


Figure 22 - Impact of barrier thickness of copper interconnect on memory access time.

3.6 Alternative Materials to Replace Copper Interconnect

More than a decade ago, the semiconductor industry switched from Cu to Al due to its superior conductivity and better resistance to electromigration. However, since the bulk electron mean free path (MFP) in Cu is much larger (40nm) than Al (16nm), the size effects are more pronounced in Cu than Al at nanoscale dimensions. This leads to a larger increase in the copper resistivity, and copper loses its conductivity advantage over Aluminum. Moreover, Al wires potentially may not require a diffusion barrier, which leads to larger effective cross-section and lower resistance for Al wires [33]. Due to these trade-offs between Cu and Al in nanoscale dimensions, we construct models to

investigate a few alternative technologies and investigate their impacts on the memory performance. We compare four different interconnect technology options, where local and global wires are either Cu or Al wires. Since the number of memory banks significantly affects the memory access time, we study and compare the impact of using various interconnect technology options for a memory system with a low number of banks (64), and a high number of memory banks (4096) as shown in Figure 23. In a memory system with a small number of banks, the local interconnects; i.e. the wordline and bitline, dominate the memory delay. Therefore, replacing Cu with Al in the local interconnects results in the highest delay reduction. In a memory system with a large number of banks, the main contribution to the memory delay comes from the global interconnects; i.e. the address and data interconnects. Therefore, to reach the highest delay reduction, Al should substitute Cu in the global interconnects.

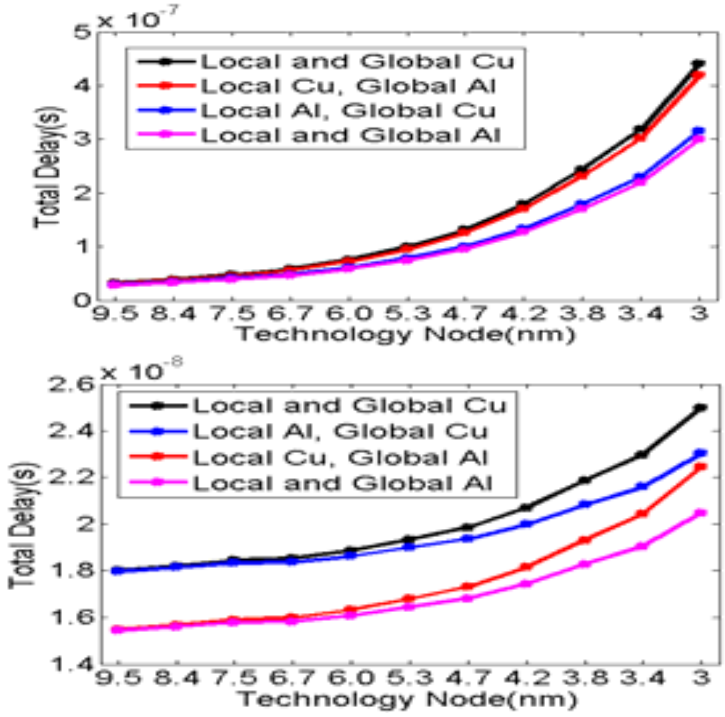


Figure 23 - Impact of alternative hybrid Al-Cu interconnect technologies on memory system access time for future technology generations for memory systems with (top) 64 banks and (bottom) 4096 banks.

3.7 Memory Performance and Bottlenecks through Scaling

We study the delay and dynamic power for various technology generations from 9.5nm down to 3nm, which corresponds to the time span of the ITRS years of 2016 to 2026. Two memory systems are investigated, which include 64 and 4096 banks, respectively as shown in Figure 24. In a memory system with a small number of banks, the wordline and bitline dominate the memory delay. In addition, the memory access time changes by a large percentage in the future technology nodes because their length changes considerably with the scaling. In a memory system with a large number of banks, the main contribution to the memory delay comes from the global address and data interconnects. Since the main parameters that affect the delay of global address and data interconnects are the permitted added wiring area and the number of banks, their delay and hence the memory access time do not change much with the scaling.

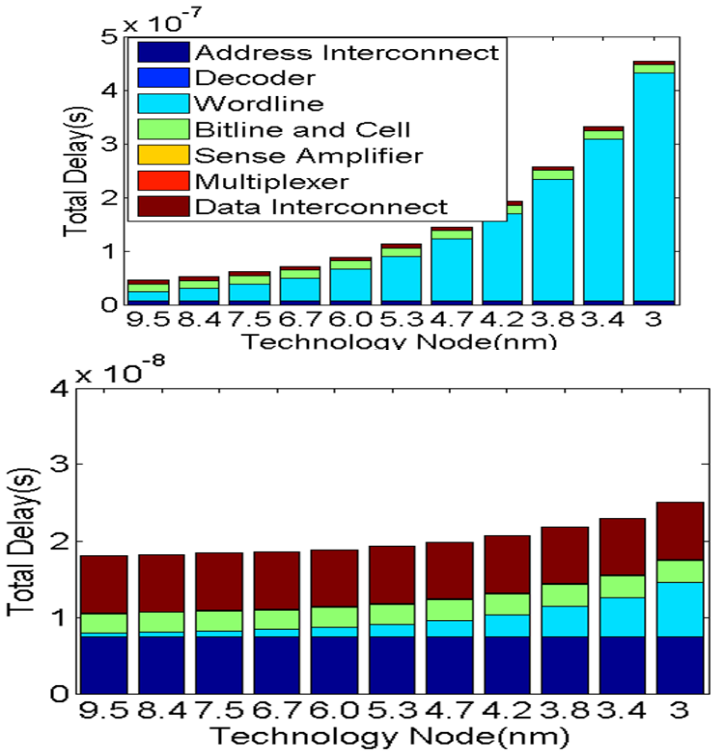


Figure 24 - Memory system access time through scaling in future technology generations for a memory system with (top) 64 banks and (bottom) 4096 banks.

Figure 25 shows the dynamic power dissipation for a memory system with 64 and 4096 banks for various technology generations. For any number of banks, the main sources of system dynamic power dissipation are the address and data interconnects.

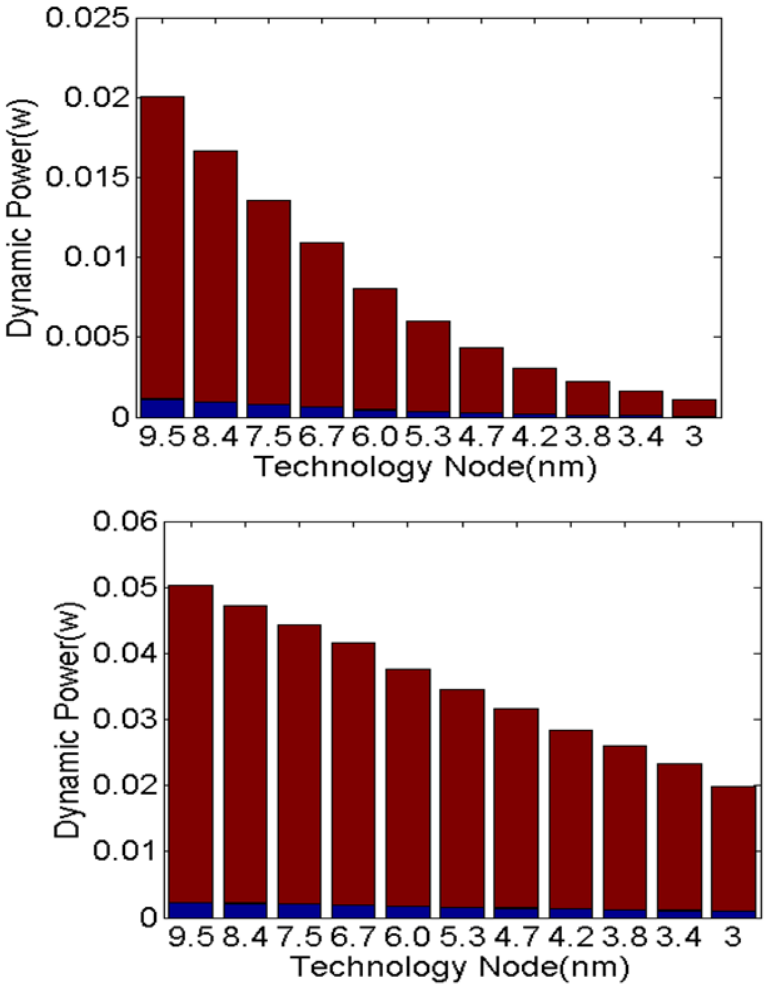


Figure 25 - Memory system power through scaling in future technology generations for a memory system with (top) 64 banks and (bottom) 4096 banks.

CHAPTER 4 3D MEMORY ARRAYS

4.1 Introduction

The DRAM industry has continued to innovate both technologies and architectures in order to improve the performance, power, capacity, and cost of DRAMs. New materials and fabrication processes have been steadily introduced. To address the challenges associated with DRAMs performance, cost and scaling, both technology and circuit solutions should be investigated. One of the most promising solutions is 3D memory integration.

The popularity of 3D Stacked ICs (3D-SICs) is rising among industry and research groups [34-37]. 3D-SIC technology based on TSVs provides numerous advantages as compared to traditional 2D-ICs, emerging as one of the main competitors to sustain Moore's Law [1]. Stacking dies with vertical interconnects offer many benefits [34], such as (a) low latency between adjacent dies, (b) reduced power consumption, (c) high bandwidth communication, (d) improved form factor and package volume density, and (e) heterogeneous integration. One of the main applications that utilize the mentioned benefits is the stacking of memory dies.

There have been many publications on the interconnect scaling issues in logic chips and how 3D integration can potentially address some of these issues [30,31,38-41]. Stacking memory on logic has also been the subject of extensive research [42-43]. However, there has been no comprehensive study on the performance and scalability of interconnects in memory arrays and how 3D integration can potentially address some of these challenges which is the subject of this chapter. While interconnects have induced

many challenges for the integrated circuit technology in the past decades, there have been major changes in the nature and the severity of the challenges in recent years. In the past, only the long global interconnects imposed limits on the chip clock frequency since the delay of local interconnects scaled with technology. However, the increase in the copper resistivity due to size effects, such as surface/grain boundary scattering, and line edge roughness, has led to a significant increase in local interconnect delay causing it to become a challenge [44-45]. Size effects are particularly problematic for memory arrays as bitlines and wordlines have the tightest pitch to minimize the cell area.

In Section 4.2, three flavors of 3D memory integration are considered and the memory access time, die area, and dynamic power are quantified for each option. Section 4.3 describes the main challenges of DRAM scaling and potential solutions offered by 3D integration. Section 4.4 studies the impact of TSV and MIV technologies on 3D memory performance. Section 4.5 studies the scaling trends for 3D memory at various technology generations from 9.5nm down to 3nm and quantifies the benefit of the transition from 2D to 3D memory for different memory array architectures. Finally, Section 4.6 presents the conclusion.

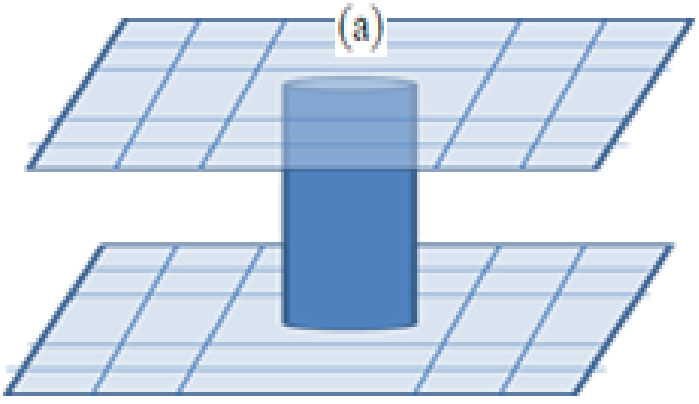
4.2 Three-Dimensional DRAM Chips

In this section, three flavors of 3D integration are considered and the memory access time, memory die footprint area, and dynamic power are quantified for each option. In all these options, the total capacity of the die-stacked memory chip is constant, and the footprint area of the memory array decreases as the number of stacked dies increases. The values of the total memory capacity and the total number of memory banks are chosen as 7.3GB and 218, respectively. However, similar trends can be

observed for other values of memory capacity and bank number as well. The values of TSV parameters such as resistance and capacitance are found using the models in [46].

4.2.1 3D Integration at the Memory Array Level

In this option, a bundle of TSVs at the center of the chip connects various stacked dies together. The memory banks at each die are two-dimensional whereas the dies are stacked in the third dimension. The global address and data interconnects are transmitted to the upper dies through a bundle of uniform bus TSVs located at the center of the dies. For the local interconnects; i.e. the wordlines and bitlines, on-die wiring is used. Each TSV in the TSV bundle has the diameter of 20μm, an aspect ratio of 5, and a pitch of twice its diameter. As seen in Figure , this 3D memory structure with 4 stacked dies results in 27% memory access time reduction. Also, the added area due to TSVs is negligible compared to the memory banks area.



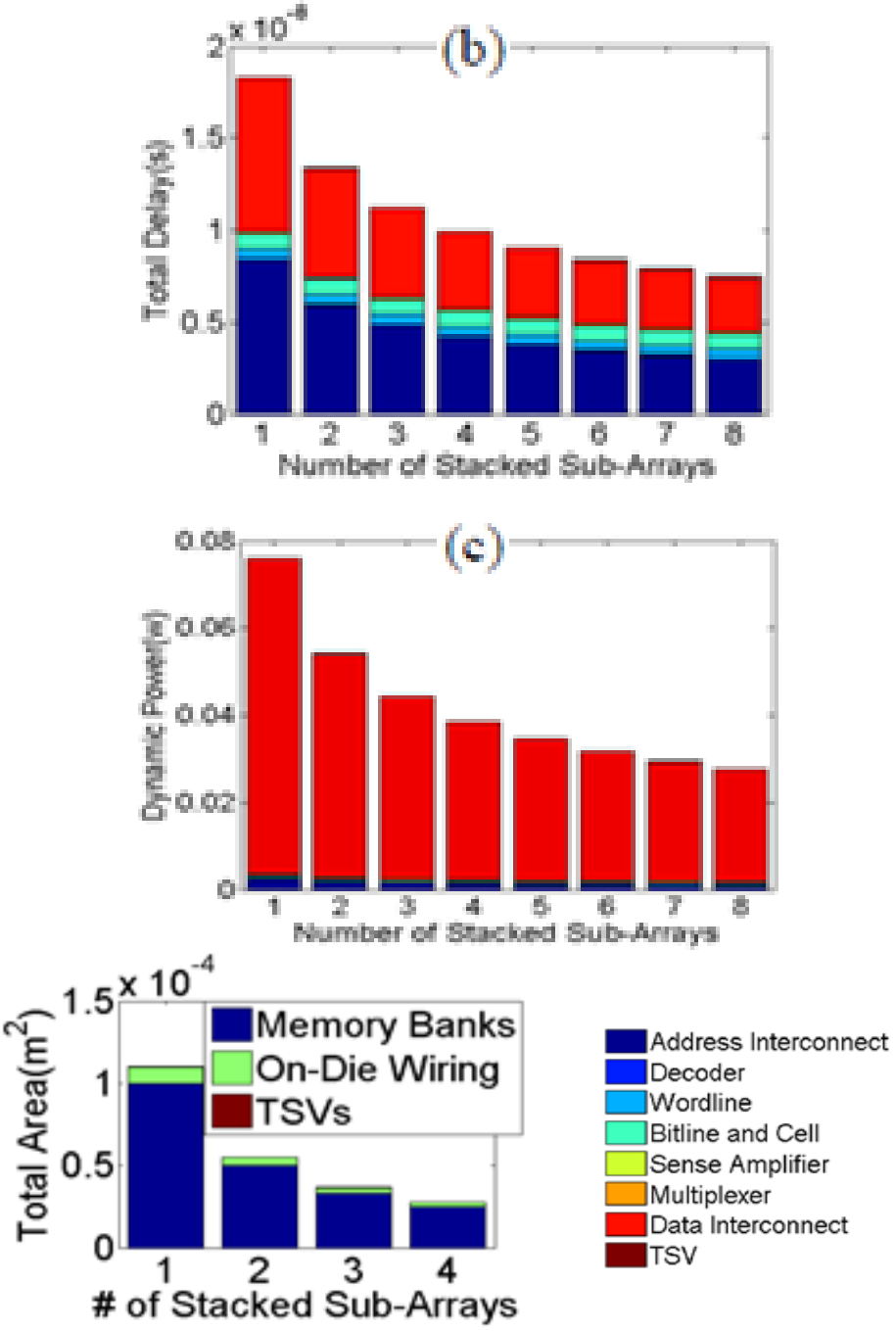


Figure 26 - (a) The memory structure, (b) access time, (c) dynamic power, and (d) area of the 3D array level memory structure.

4.2.2 3D Integration at the Memory Bank Level

In the 3D memory configuration studied in the previous subsection, the potentials of 3D integration are not fully exploited. This is because the memory banks at each die are still two-dimensional. Alternatively, 3D integration can be utilized at a lower level in memory and the memory banks can be formed as 3D memory segments. In this option, the memory array is an array of 3D sub-array memory banks. The address and data interconnects for each bank are transmitted to the upper stacked banks through TSV bundles located at the center of each memory bank. For the local interconnects; i.e. the wordlines and bitlines, on-die wiring is used. As seen in Figure , the three-dimensional memory at memory bank level reduces the memory access time by 37% which is more than the 27% memory delay reduction achieved by the three-dimensional memory at memory array level investigated in the previous subsection. This is due to the address and data being transmitted to the upper banks through the TSVs instead of being transmitted to the die center through the TSVs and from the die center to the memory banks through the on-die wires. However, as seen in Figure , there is a trade-off between memory delay reduction and increase in memory die footprint area. In this option, since each memory bank has its own bundle of TSVs, the added area due to TSVs is much greater than in the 3D memory at memory array level. The dymamic power also increases significantly as the capacitance associated with the TSVs is substnatilly larger than that of on-chip wires.

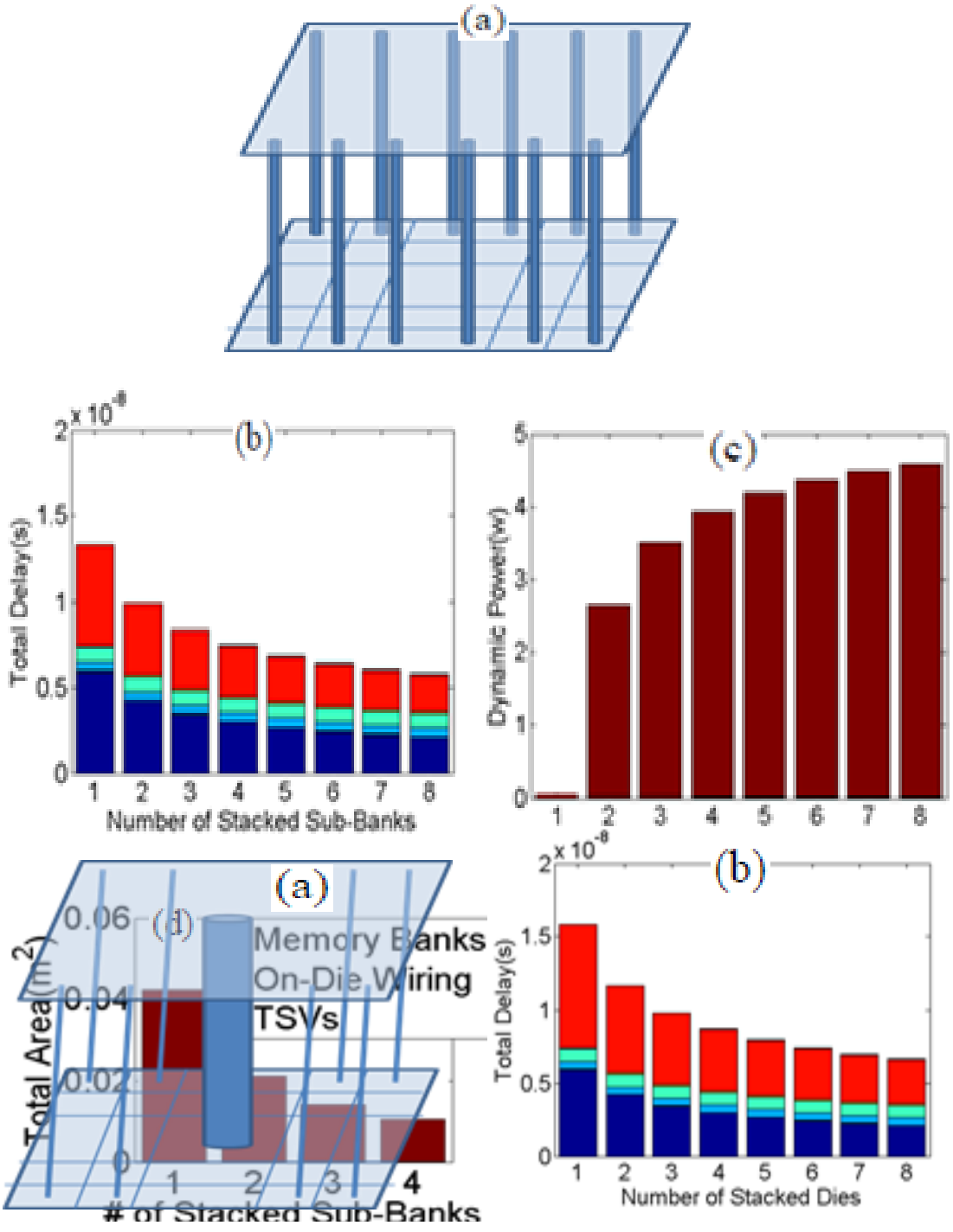
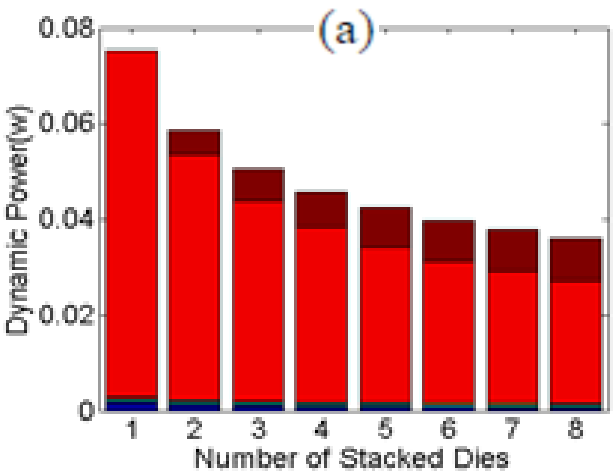


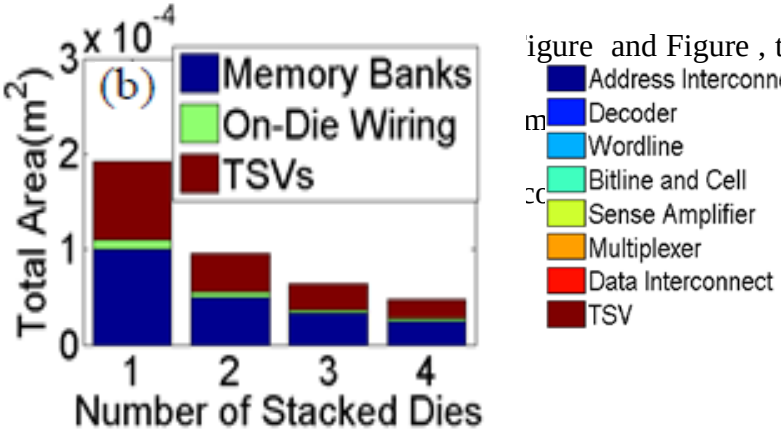
Figure 28 – (a) The memory structure, and (b) the access time of the optimized 3D memory structure.

d)

subsections 4.2.1 and 4.2.2, an optimized configuration for 3D die-stacked memory array

is presented i its features from the 3D memory array 3D memory bank level integration. I sub-arrays and the 3D memory is t interconnects; i.e. the wordlines an dress interconencts are trasmitted to d at the center of each bank. The gl dies using TSV bundle loca igure and Figure , this option is the optimal 3D m while maintaining less add c) bank level memory stru





4.3 Impact on Required Cell Storage Capacitance

Architecture in DRAM could limit further more charge on the capacitor, higher voltage levels are applied to the cell transistor gate and longer channel lengths are used in DRAM compared to high performance logic chips while having the same lithography capability [47]. In the past few technology generations, the bitline capacitance has stayed around 25-40fF [48]. Therefore, in order for the signal voltage to stay above the minimum threshold value of the sense amplifier, the charge capacity of the cell storage capacitor should be maintained while scaling down its size. A possible way to achieve this is to increase the applied voltage levels to the capacitor. However, as DRAMs are scaled to smaller dimensions, the voltage applied to the gates is required to follow a scaling path similar to the one for the voltage levels in logic chips. At

Figure 29 - (a) dynamic power, and (b) area of the optimized 3D memory structure.

each technology generation, the field strength at DRAM devices is at the maximum value allowed for gate-oxide reliability [47]. As a result, the need for keeping the charge storing capability of the cell capacitor at a constant value while scaling down its size has become a roadblock in the path of scaling down the DRAM device.

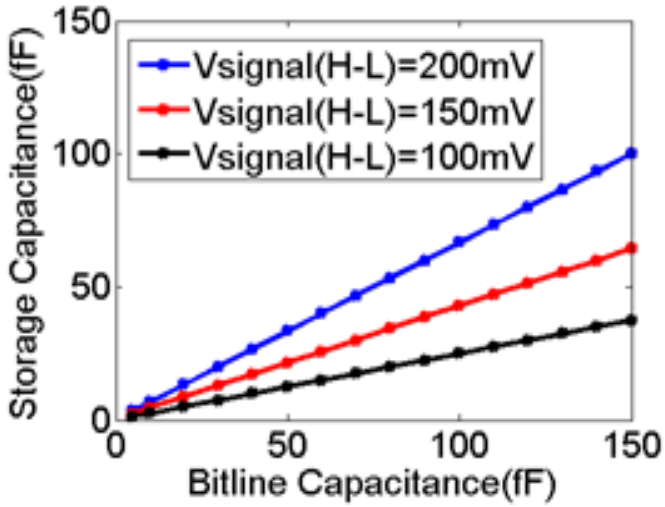


Figure 30 - The required value for storage capacitor for different values of bitline capacitance and minimum required voltage swing at the sense amplifier input in the optimized 3D memory.

In the 3 the dyanmic power consumption ; the memory die into a few stacked and 2) using TSVs to replace some emory arrays with a much larger n in shorter length and smaller capac lt of smaller bitline capacitance, t maintaining a constant input voltage onfiguration resolves the challenge cell capacitor while scaling down , the required cell capacitor is scaled down by reducing the bitline capacitance. By scaling down the cell capacitor, the bitline and cell delay, the memory access time, and the cell refresh power are greatly reduced, as shown in Figure , Figure , and Figure . Figure and Figure show the values of the bitline capacitance and the storage capacitance for the 2D and 3D memory systems. As the memory allowed power consumption increases, the number of memory banks could be increased which leads to smaller values of the bitline capacitance and the storage capacitance.

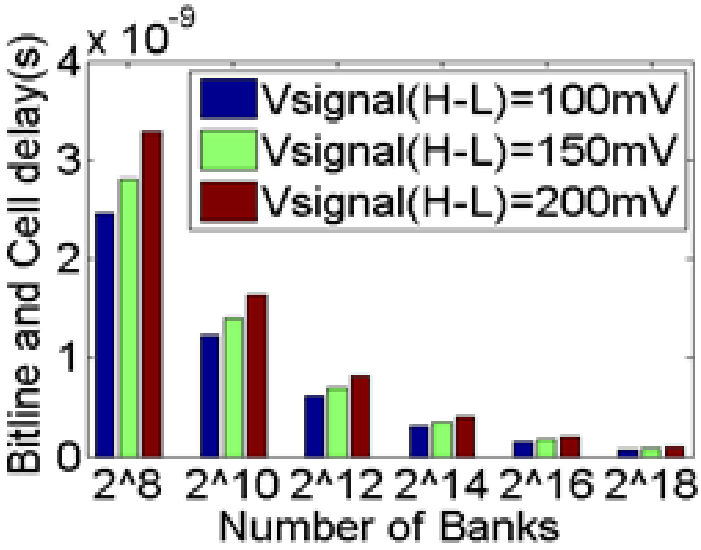


Figure 32 - Impact of minimum required voltage swing at sense amplifier input (and hence the cell capacitor value) on bitline and cell delay for different numbers of memory banks in 3D memory.

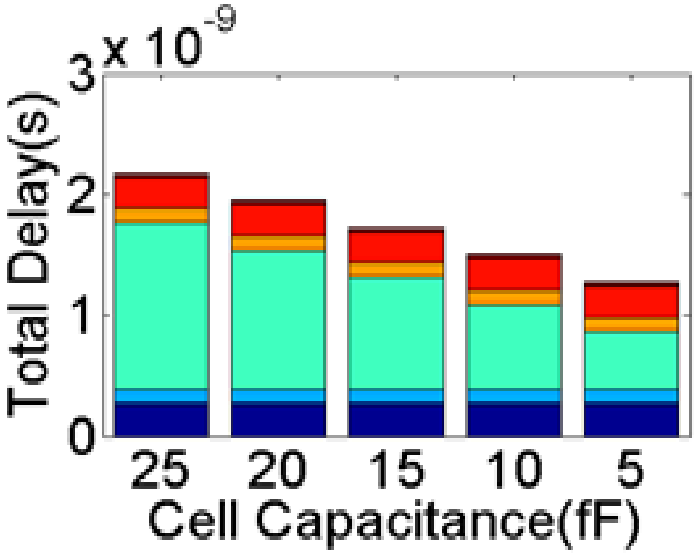
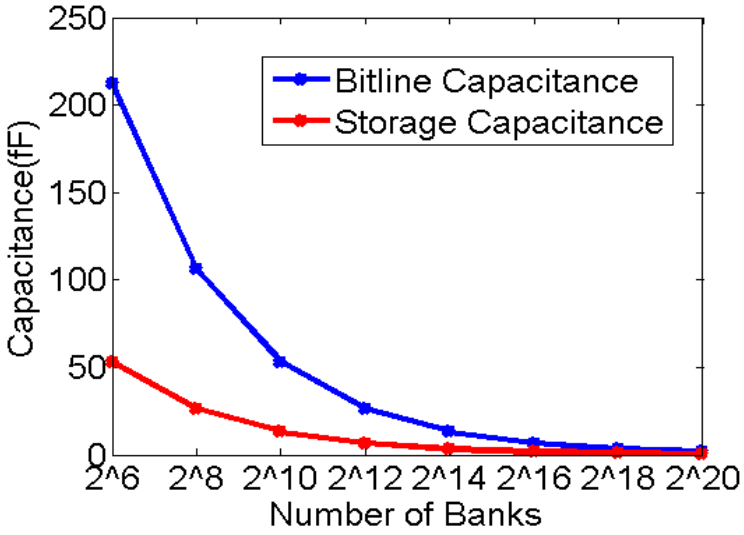


Figure 31 - Impact of cell capacitor value on memory access time in the optimized 3D memory.



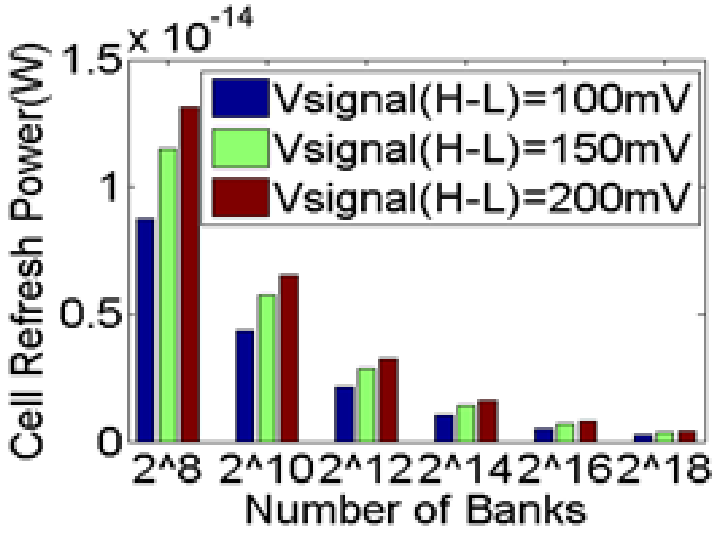


Figure 33 - Impact of minimum required voltage swing at sense amplifier input (and hence the cell capacitor value) on cell refresh power for different numbers of memory banks in 3D memory.

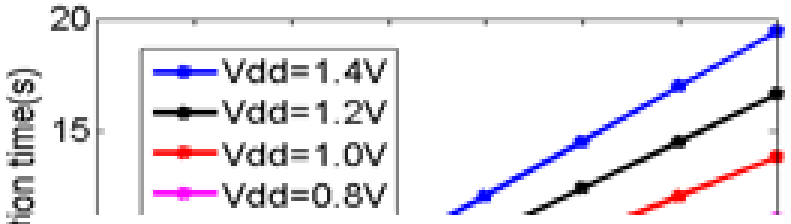


Figure 34 - Values of bitline and storage capacitance with the sense amplifier input voltage swing of 100mV for different numbers of memory banks in 3D memory.

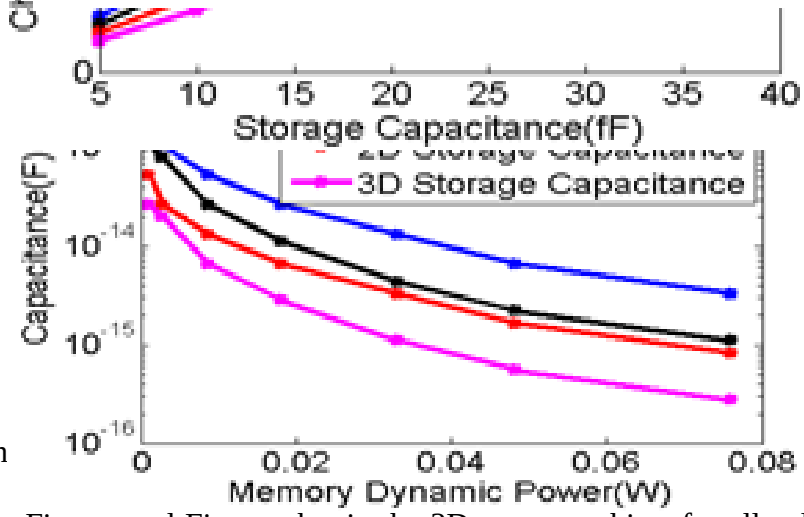
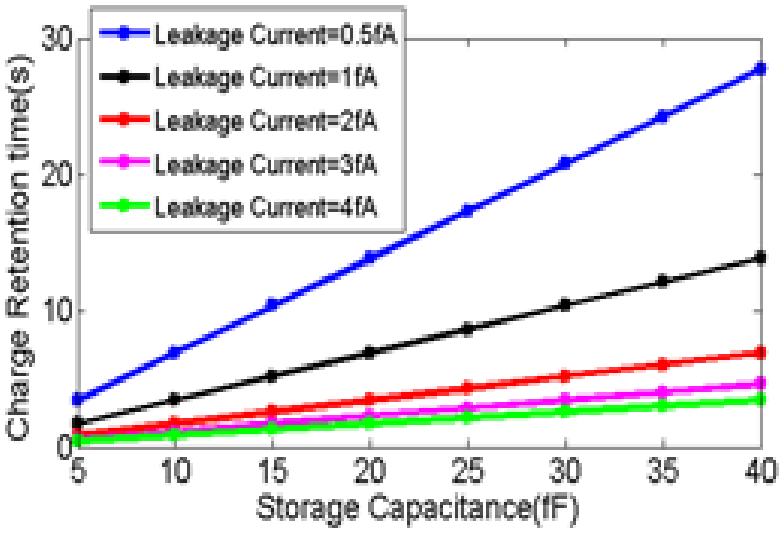


Figure 35 - Values of bitline and storage capacitance with the sense amplifier input voltage swing of 100mV for different values of memory dynamic power in 2D and the optimized 3D memory structure.

An ention time. It is shown ir s of V_{dd} and the leakage till an order of magnitu ime in DRAMs which is er, reducing the cell storage capacitor value does not lead to the deterioration of the memory array performance.

Figure 36 - Impact of cell storage capacitor value on cell capacitor charge retention time for different values of supply voltage.



4.4 3

ect the stacked

transis SV and MIV

techn are studied. An

impor n in FIGURE ,

Figure 37 - Impact of cell storage capacitor value on cell capacitor charge retention time for different values of leakage current.

the re ratio, the via
resista put resistance
which mall effect on
increa duction has a
strong liameter from
20µm hip with four
stacke fabrication at
small vias.

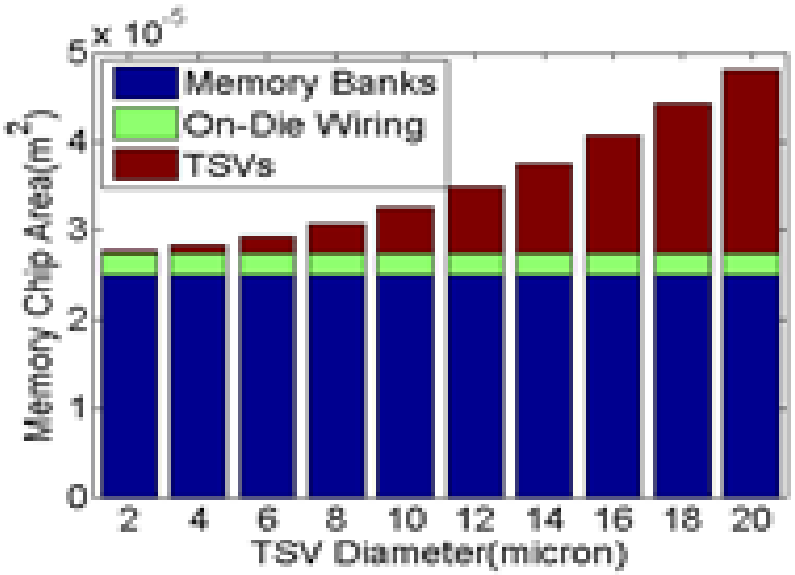


Figure 39 - Impact of TSV diameter on memory array area in the optimized 3D memory system.

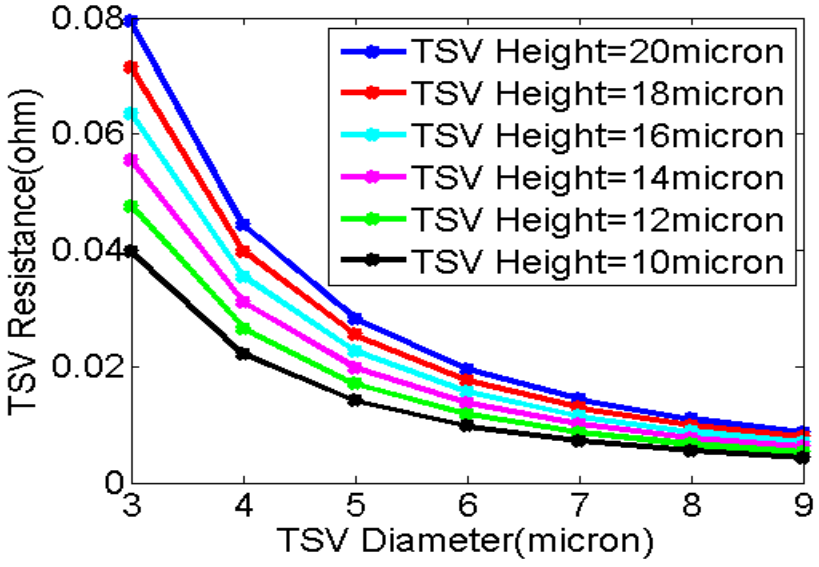


Figure 38 - Impact of TSV diameter on (a) TSV resistance for different values of TSV height in the optimized 3D memory system.

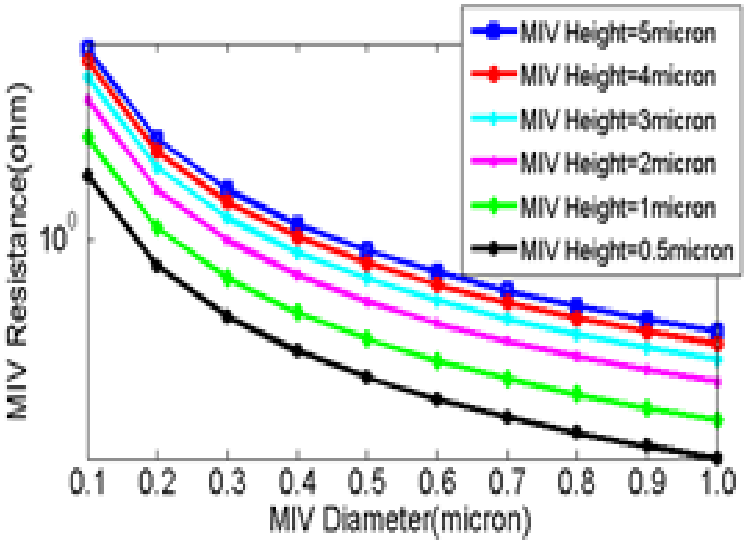


Figure 40 - Impact MIV diameter on MIV resistance for different values of MIV height.

An al -tier via (MIV). Using the n. Figure and Figure sho e and via delay for differer neters down to 100nm and via heights up to 5µm, the via resistance is at least two orders of magnitude smaller than the driver resistance which is commonly around 1-10kΩ. Therefore, as shown in Figure , for these ranges of via diameter and via height, the via delay is five orders of magnitude smaller than the memory access time which is around 10^{-8}s at 9.5nm technology node, as shown in **Error! Reference source not found.**. The small value of the MIV delay even with small diameters is due to its ultra short height compared to the length of the on-die wires. As a result, the MIV diameter could be reduced to 100nm without deteriorating the memory access time and frequency.

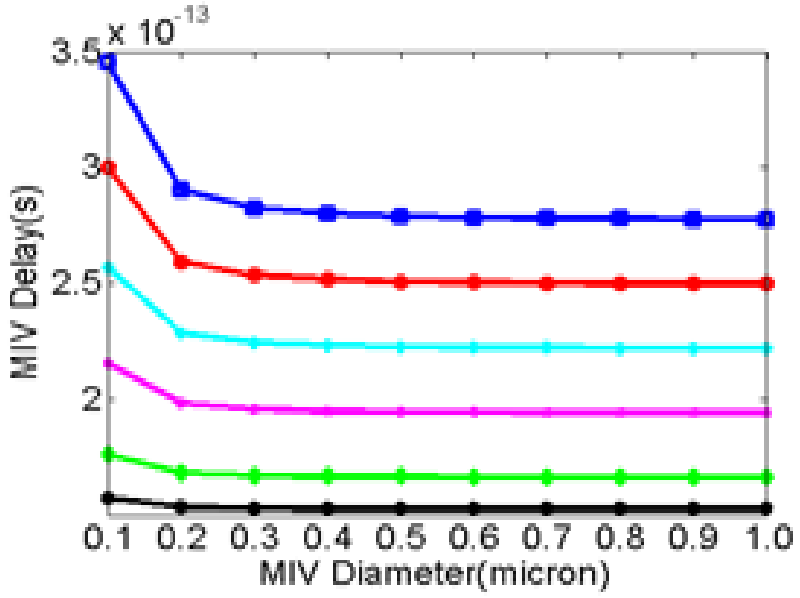


Figure 41 - Impact MIV diameter on MIV delay for different values of MIV height.

4.5 Scal

We erations from 9.5nm do 2016 to 2026. Two mem respectively.

As shown in Figure 42, in a memory system with a small number of banks, the wordline and bitline dominate the memory delay. In addition, the memory access time changes by a large percentage in the future technology nodes because their length changes considerably with the scaling. In a memory system with a large number of banks, the

main contribution to the memory delay comes from the address and data interconnects, as

seen y of address and data

inter elay

and ion,

sinc the

addi n in

tran increasing the number

of m

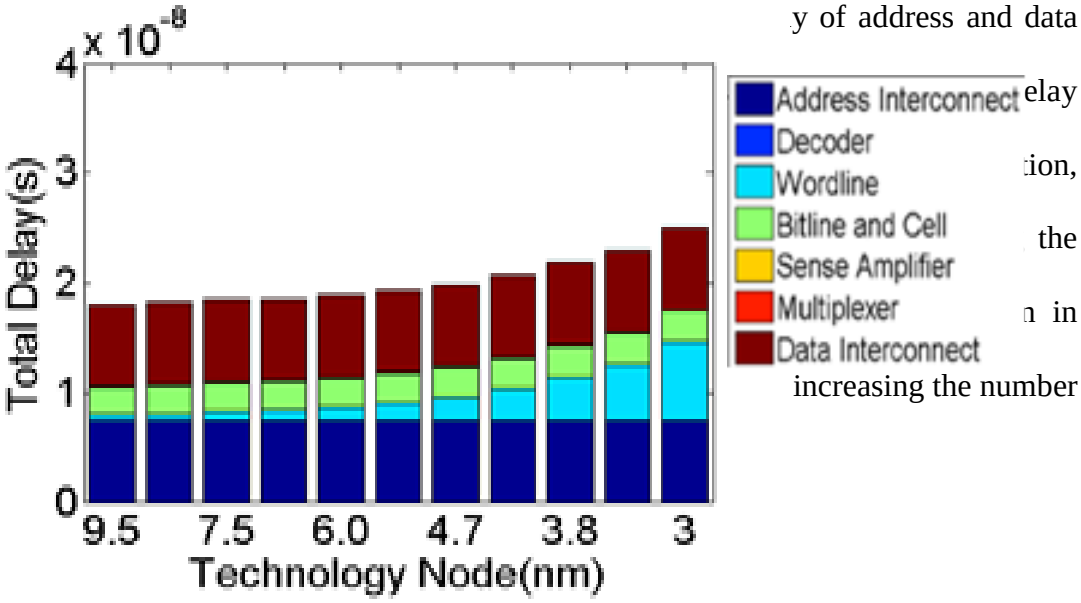


Figure 27 - Memory system access time through scaling in future technology generations for a memory system with 4096 banks in a 2D memory structure.

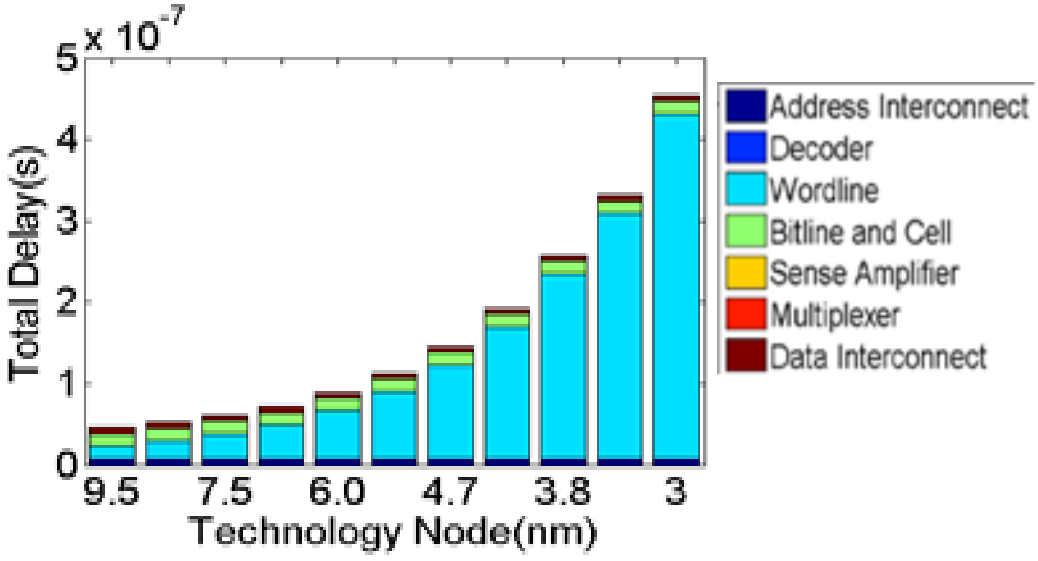


Figure 26 - Memory system access time components through scaling in future technology generations for a memory system with 64 banks in 2D memory structure.

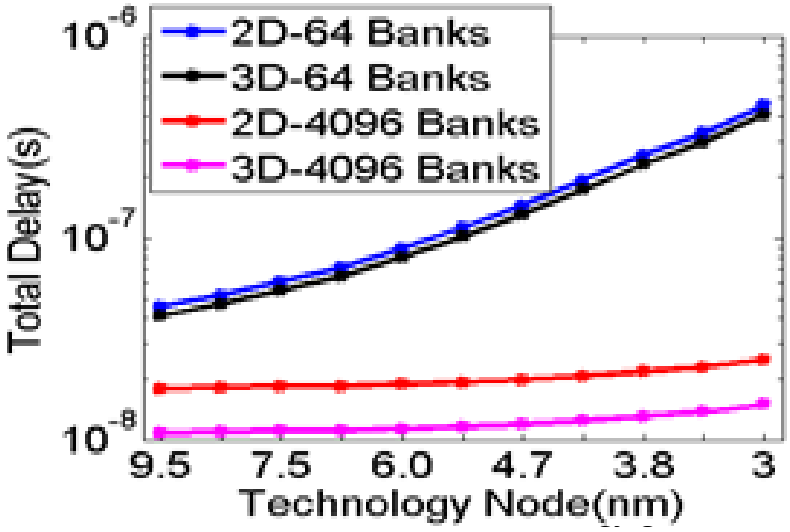


Figure 28 - Memory system access time through scaling in future technology generations for a memory system with 64 or 4096 banks in 2D or 3D memory structure.

F n for memory systems any number of banks, interconnects. As a re e 2D to the 3D memory system does not change with increasing the number of memory banks, as shown in **FIGURE 47**.

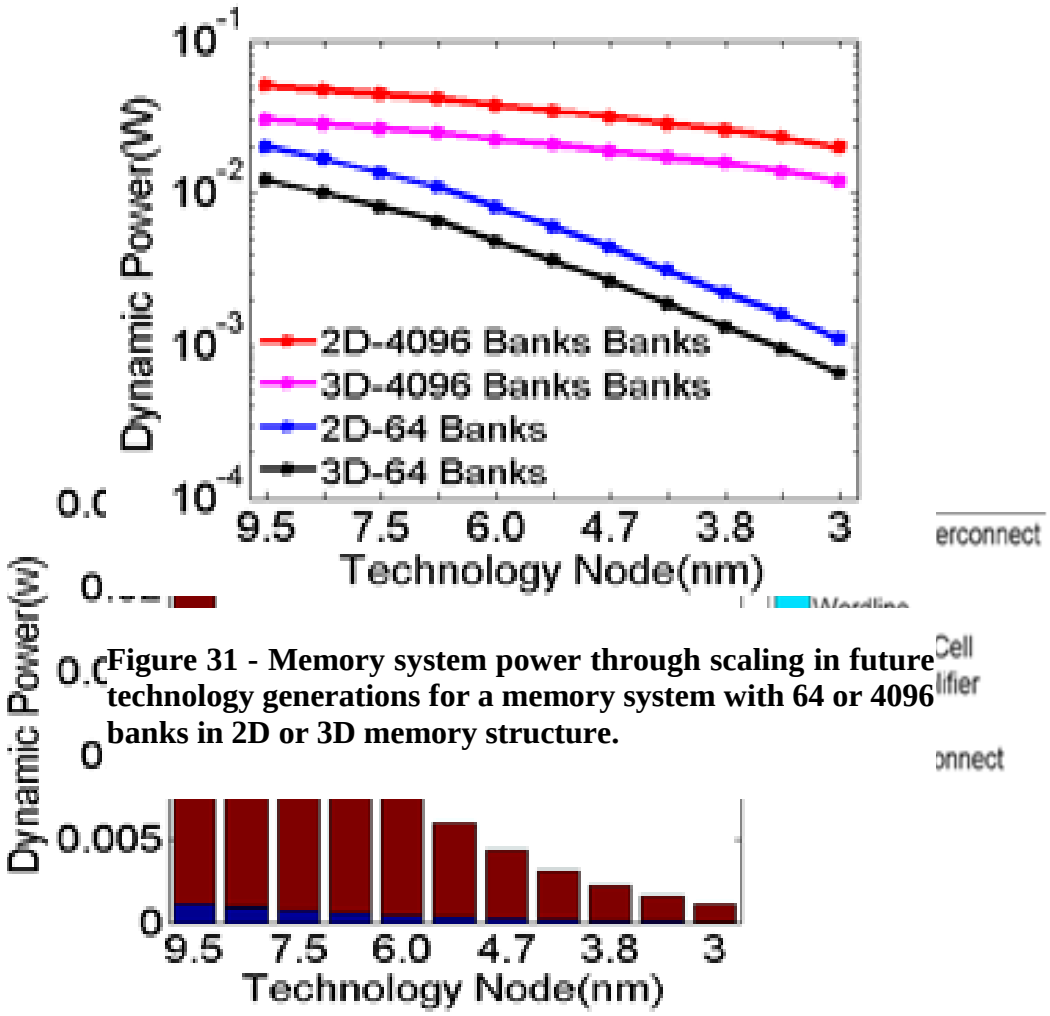


Figure 31 - Memory system power through scaling in future technology generations for a memory system with 64 or 4096 banks in 2D or 3D memory structure.

Figure 29 - Memory system power through scaling in future technology generations for a memory system with 64 banks in a 2D memory structure.

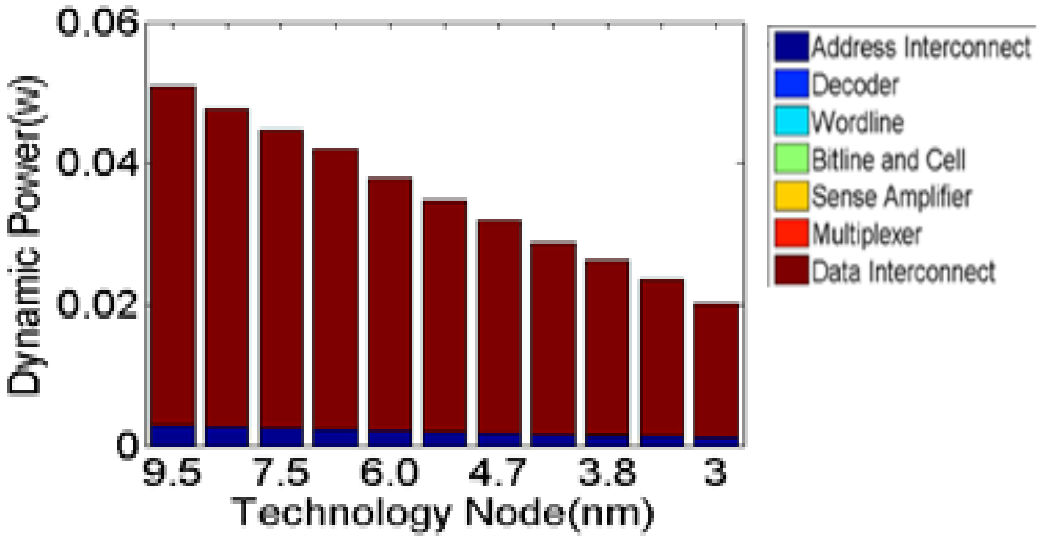


Figure 30 - Memory system power through scaling in future technology generations for a memory system with 4096 banks in a 2D memory structure.

Chapter 5 SPIN-TRANSFER TORQUE MAGNETIC RANDOM ACCESS MEMORY ARRAYS (STT-MRAM)

5.1 Introduction

As the CMOS technology advances to the deep nanoscale era, DRAM scaling faces serious challenges in speed, bandwidth, capacity, and cost. For example, historically, CPU performance has improved at an annual rate of 55% while the memory access time has only improved by 10%, resulting in the well-known memory wall problem [49]. Moreover, DRAM capacity had increased 4 times every 3 years for decades, but is now scaling more slowly, resulting in a memory capacity wall problem. Power and cost of DRAMs are also facing similar challenges [49].

Emerging non-volatile memory technologies are being investigated as potential solutions, and STT-MRAM is one of the promising technologies among them. Remarkable progresses in STT switching with MgO magnetic tunnel junctions (MTJs) and increasing interest in STT-MRAM in the semiconductor industry have been witnessed in recent years. A key milestone in STT research has been reached in early 2004 by first demonstration of STT switching in Al_2O_3 based MTJs by Huai et al. [50-51]. Subsequent STT research has been focusing on MgO MTJs. And in 2005, STT switching has been successfully demonstrated in MgO MTJs with TMR> 150% and small intrinsic switching current density $Jc0 = 2 – 3 \times 10^6$ A/cm^2 [52-53].

Until now, there has not been a comprehensive analytical model investigating the STT-MRAM interconnect performance bottlenecks and reliability which are the focus of this chapter. The interconnect reliability challenges and the performance-reliability tradeoffs are also studied in this chapter and interconnect optimization techniques are

introduced to reduce the critical interconnect delays. Finally, an alternative memory cell structure and its impact on the memory performance are investigated.

5.2 Model Approaches and Assumptions

Table 3 shows the important parameters of the STT-MRAM model.

Table 3 - Important parameters of the STT-MRAM model.

Parameter	Value	Ref.	Parameter	Value	Ref.
Chip size	10mm×10mm		MTJ R_H	50Ω	
Cell architecture	Reverse 1T/1MTJ	[54]	MTJ TMR	150%	
Cell size	40 λ^2	[55]	AlO_x MTJ J_{c0}	0.14mA	[56]
FinFET(L/W)	20/135nm	[55]	MgO MTJ J_{c0}	0.04mA	[56]
V_{dd} (core/IO)	0.8/1 V	[55]	wire AR	1.6	[24]
MTJ size	30/75nm		Standby current	0	[57]
MTJ R_L	20kΩ				

5.2.1 Memory Cell Structure

The current density required for switching MTJs from the parallel state to the anti-parallel state is 20-50% larger than the reverse [54]. Figure shows that by reverse connecting the access transistor and the MTJ instead of the conventional way; i.e. connecting the access transistor to the free magnet instead of the fixed magnet in an MTJ, a larger value for the V_{GS} of the transistor is obtained, resulting in a larger current.

5.2.2 Memory Subarray Architecture

The memory array is divided into a number of banks. The global interconnects transmit address and data from the array input to the bank, forming a hierarchical-tree (h-tree) network. The wordlines (WL) are connected to the gates of the transistors, the

sourcelines (SL) are connected to the source of the transistors, and the bitlines (BL) are connected to the MTJ fixed electrodes. Elmore delay model is used to calculate the interconnect delay. An optimum number of repeaters are placed along the global interconnects. For the wordlines and bitlines, no repeaters are used due to the cell size limit. To calculate the delay of the wordlines, the capacitance of transistors connected to the wordlines is added to the wire capacitance. Figure shows the memory subarray structure, and Figure shows the memory cross-section view [55-56].

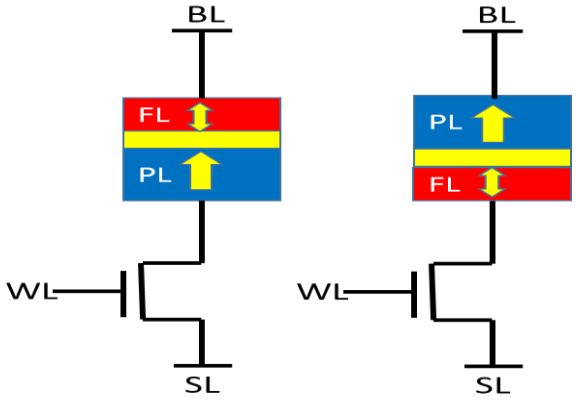


Figure 48 - Conventional (left), and reverse-connected (right) MTJ cell structures.

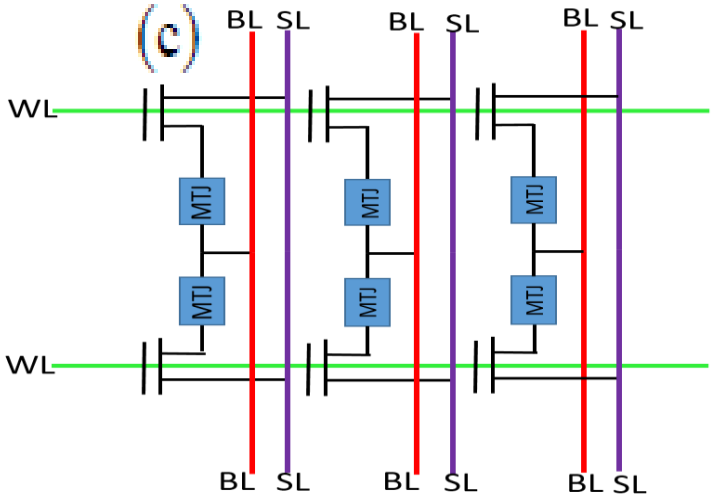


Figure 49 - STT-MRAM subarray structure.

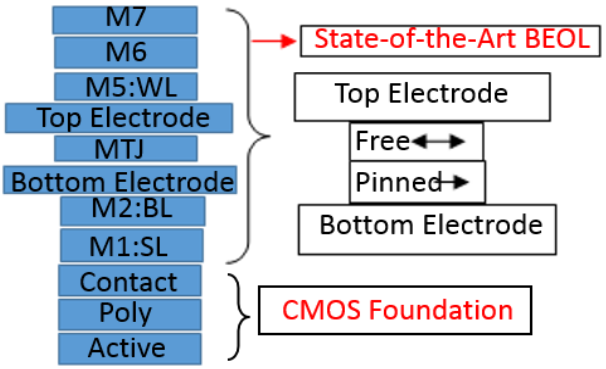


Figure 50 - STT-MRAM chip cross-section view [7-8].

For writing a "0" value to a memory cell, a high voltage is put on the WL, the BL is connected to V_{DD}, and the SL is connected to GND, which results in MTJ switching from the low-resistance (LR) state to the high-resistance state (HR). For writing a "1" to a memory cell, the BL and SL voltages are reversed, which causes the MJT to switch from HR to LR. The reading process is similar to writing a "0" value, but I_{Read} is much smaller than I_{Write} in order to avoid flipping the MTJ while reading the cell value.

There are different directions in the field of STT-MRAM research that include optimization of MTJ, cell structure, memory configuration, layout, and interconnects. To find the most crucial research direction, the results for the memory performance are studied in order to find the bottlenecks of the STT-MRAM performance. The numbers are for a memory array with 2^8 memory banks. However, the insights obtained regarding the limits and opportunities associated with interconnects apply to memory arrays with any numbers of banks.

5.3 Model Results and Discussions

Figure 51 shows the total memory delay. The delay of peripheral circuits such as decoders, multiplexers, and amplifiers have been included in the model, but their values are negligible and not visible in the figure. The main contributors to the memory delay are the local and global interconnects, and the cell. For a small number of banks, since the banks are large and the WLs and the BLs are long, most of the delay comes from the local interconnects inside the banks; i.e. the WL and BL. For a large number of banks, most of the delay comes from the global address and data interconnects. As the bank number increases and the total memory delay gets smaller, the contribution of the bitline and cell delay becomes more notable. The cell delay is calculated as the average of the MTJ read and write pulse widths.

Figure 52 shows the total dynamic power consumption of the memory. Most of the dynamic power is consumed in the global address interconnects, data interconnects, and the cell. Data interconnects dissipate far more energy than the address interconnects since they transmit decoded data, and as a result require a much larger number of wires. Figure 53 shows the total access energy and EDP. Figure 54 shows the memory logic circuits area.

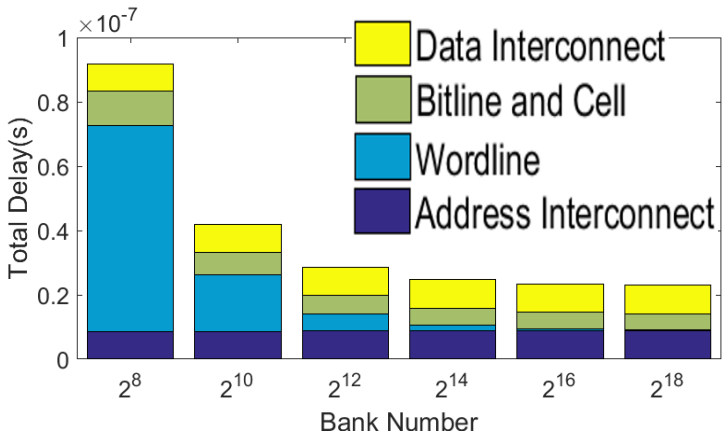


Figure 32 - STT-MRAM access time components.

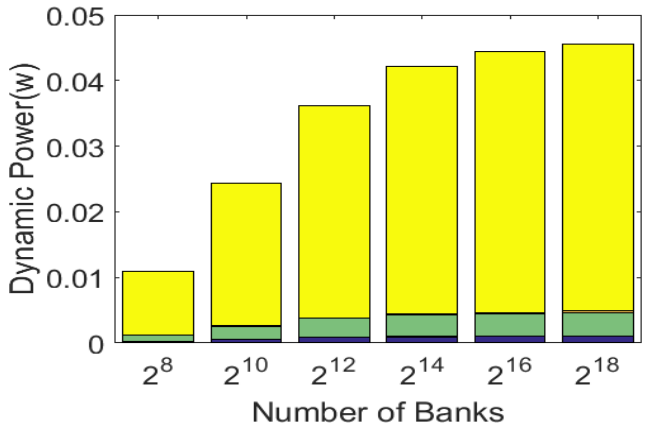


Figure 33 - STT-MRAM dynamic power consumption components.

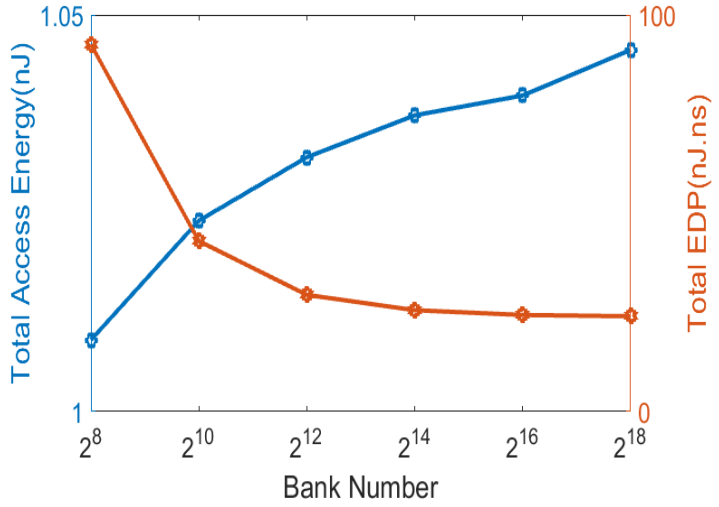


Figure 34 - STT-MRAM total access energy and EDP.

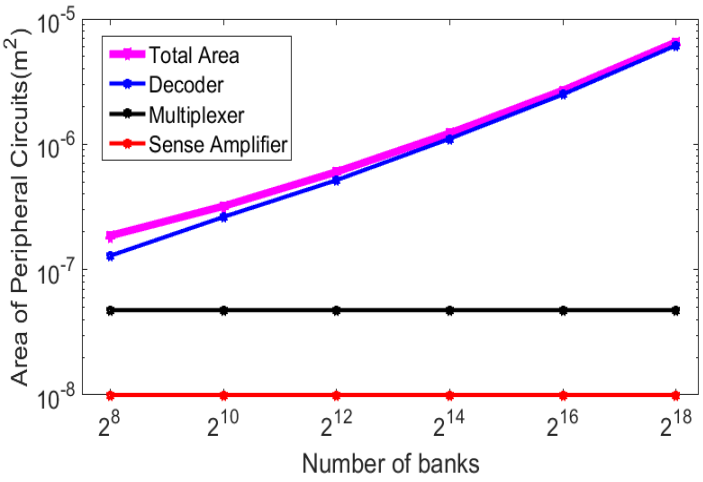


Figure 35 - STT-MRAM peripheral circuits area.

5.4 Interconnect Reliability Challenges

One of the challenges in STT-MRAM is the required high current density for switching the MTJ and the issues arising from it, such as electromigration (EM). Taking EM into account, the interconnects mean time to failure (MTTF) is found using the Black's equation [58].

$$MTTF = \left(\frac{A}{J^n}\right)\exp\left(\frac{E_a}{kT}\right)$$

where A is a constant based on the cross-sectional area of the interconnect, J is the current density, E_a is the activation energy (e.g. 0.7eV for grain boundary diffusion in aluminum), k is the Boltzmann's constant, T is the temperature in Kelvin and n is a scaling factor (with the value of 1 or 2 depending on the EM kinetics) [58].

As shown in Figure 55, the MTJ switching is divided into 3 regions, precessional switching for fast switching with pulse widths less than 3ns, slow thermal switching with

switching pulse widths more than 10 ns, and dynamic switching for intermediate switching with switching pulse widths between 3 and 10ns [59].

The STT-MRAM cell size is limited by the access transistor size capable of delivering the large required current. As a result, the wire pitch could be increased from the minimum size without increasing the number of metal levels for SL, BL, and WL. However, the number of metal levels for global interconnects will increase. Table 4 shows the impact of increasing the BL and SL wire pitch from that value by 50% on the interconnect lifetime.

Table 4 - Interconnect MTTF for different values of BL and SL pitch.

	AlO_x MTJ	MgO MTJ	AlO_x MTJ	MgO MTJ
	BL & SL pitch=60nm		BL & SL pitch=90nm	
Current Density(A/m^2)	2.36×10^{10}	6.94×10^{9}	1.05×10^{10}	3.09×10^{9}
MTTF(s)	5.76×10^{11}	1.96×10^{12}	1.29×10^{12}	4.40×10^{12}

Figure 56 shows the impact of an increase in the metal pitch on the memory latency. Figure 57 shows the interconnects MTTF versus the memory delay as the wire pitch and the cell size increase and the die area is constant. Figure shows the tradeoff between the interconnects lifetime and memory die area assuming constant memory capacity. Among all the local and global interconnects, only the BL and the SL are directly connected to the MTJ and carry the high current density required for MTJ

switching. Therefore, for improving the interconnects lifetime, only the metal pitches for the BL and the SL need to be increased.

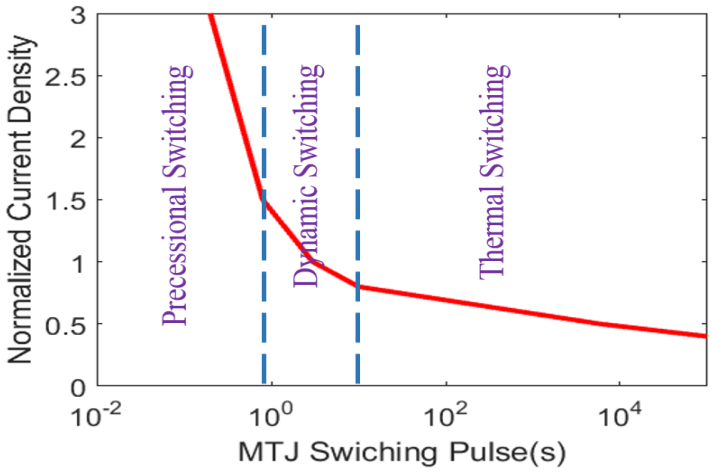


Figure 36 - MTJ switching characteristics [59].

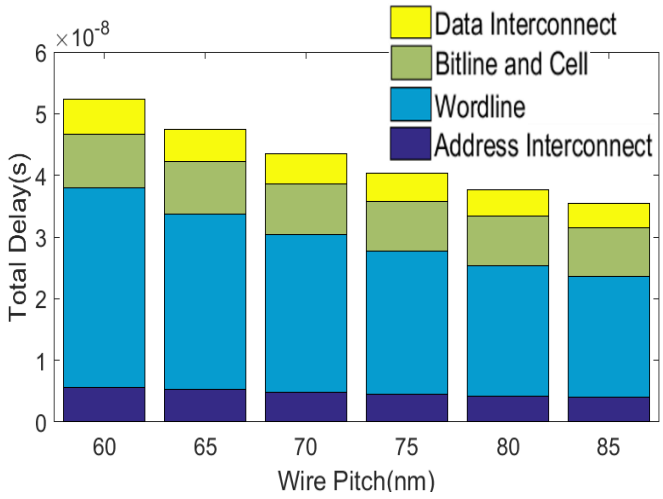


Figure 37 - STT-MRAM total memory delay as a function of wire width.

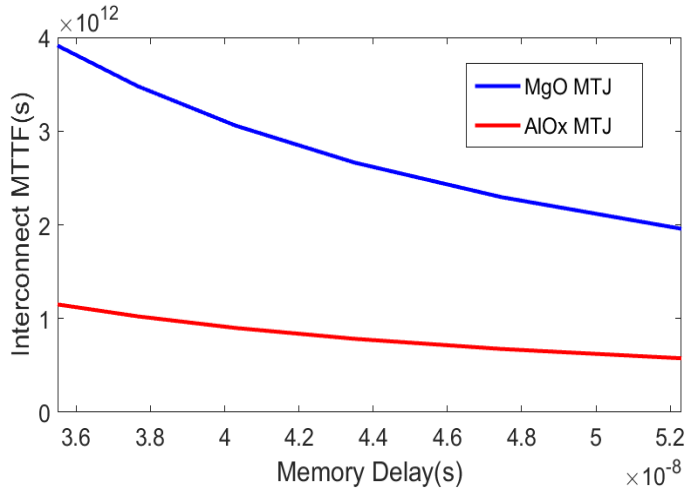


Figure 38 - Interconnect MTTF versus STT-MRAM latency assuming constant die area.

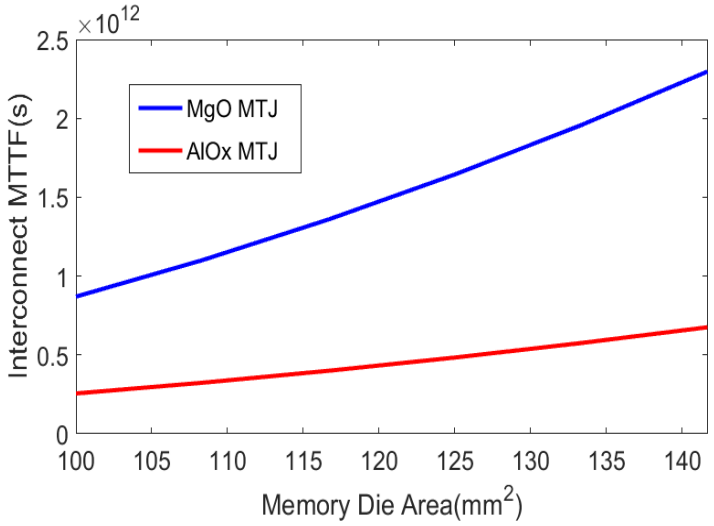


Figure 58 - Interconnect MTTF versus STT-MRAM latency assuming constant die area.

5.5 Memory Interconnect Optimization

In this section we study various wiring and memory organization schemes aimed at lowering the interconnect delay.

5.5.1 Adding Interconnect Levels

We compare the memory access times for cases where address and data interconnects are routed in different number of metal levels.

As seen in Figure , with an increase in the wire pitch and the number of metal levels, the delays of the address and data interconnects decrease. The delay of the wordline and bitline, however, remain unchanged because the wiring pitch for those wires is fixed due to the cell size limit at the 9.5nm technology node.

One of the major components of the memory access time is the wordline delay. This is because of the large resistance and capacitance associated with the wordline wire and the large capacitive load that the wordline has to drive (i.e. the transistors connected to the wordline). Conventional delay mitigation techniques such as repeater insertion and reverse scaling cannot be applied to the wordline because of the limit imposed by memory cell size.

5.5.2 Increase of Decoder Drive Current

As seen in Figure 60, by increasing the size of the decoder output drivers to four times the minimum size and quadrupling their drive current, the wordline delay is reduced by 40%. After that, the reduction is not notable as the wire resistance becomes dominant. This indicates that the role of wire resistance is much bolder in its contribution to the wordline delay than the driver resistance.

5.5.3 Optimize Bank Aspect Ratio

As seen in Figure 61, by increasing the bank aspect ratio the wordline length and delay decrease but the bitline length and delay increase. As a result, an optimal aspect ratio exists which is equal to 4:1.

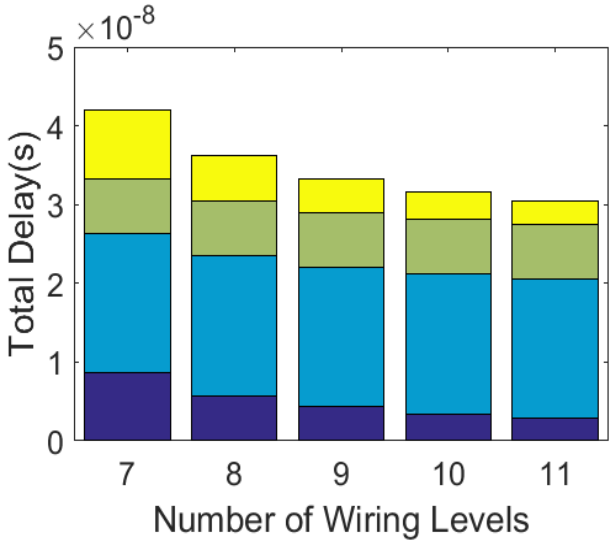


Figure 59 - Impact of the number of wiring levels on the total memory latency in STT-MRAM.

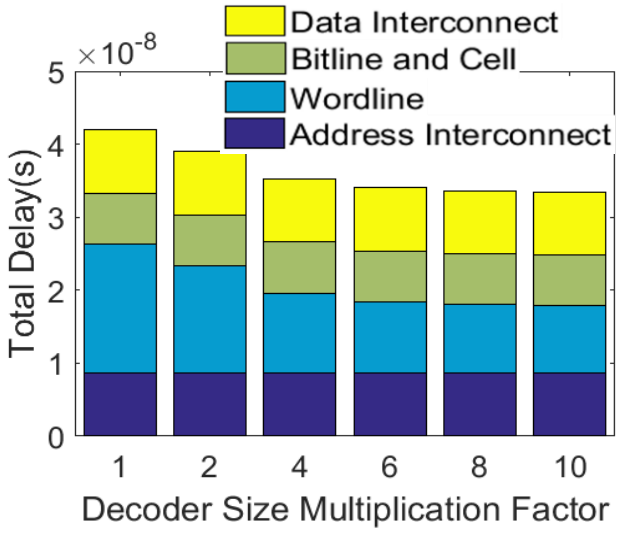


Figure 39 - Impact of decoder size on the total memory latency in STT-MRAM.

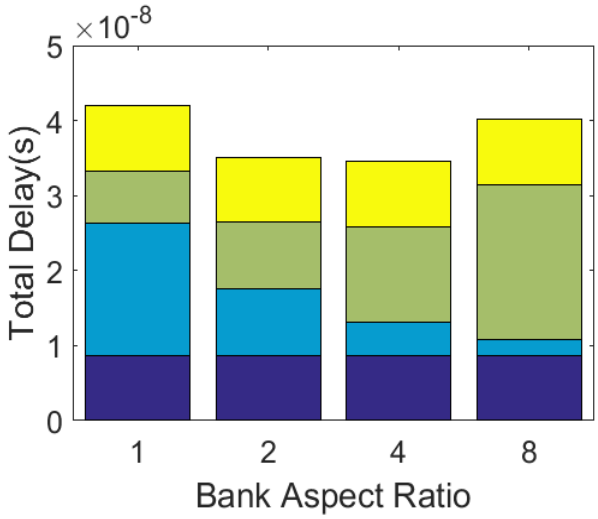


Figure 40 - Impact of memory bank aspect ratio on the total memory latency in STT-MRAM.

5.6 Potential Memory Subarray Architectures

As shown in Figure 62, a potential alternative cell architecture is the shared transistor structure [60]. Access device sharing is a potential technique for increasing memory density and reducing the demand on the cell transistor for providing large write currents with scaled-down dimensions. In the shared structure, multiple MTJs are connected to one access device, with multiple bitlines to support independent accesses. This allows the access transistor to be sized up to provide a higher write current while maintaining the same overall memory density. In addition, using one SL for multiple MTJs further contributes to the cell area reduction. A concern regarding this structure could be that during the write operation, the parasitic current paths of the MTJs connected to the access device would draw the current from the accessed MTJ, forcing the access device to be sized up, and also have the potential to flip the MTJs that are not being accessed. However, since the transistor resistance (around 5 kΩ) is much smaller

than the MTJ resistance (20-50 kΩ), the parasitic currents drawn by the other MTJs are small and not problematic.

Figure 63 shows the impact of the multi-MTJ memory cell on the access transistor size and resistance. Figure 64 shows the impact of cell MTJ number on the number of memory cells and the total memory capacity assuming that the total chip size is constant at 100mm^2. Figure 65 shows the result of this multi-bit cell on the memory delay if the total memory capacity is kept constant at 11.6Gbit and consequently the die area is reduced.

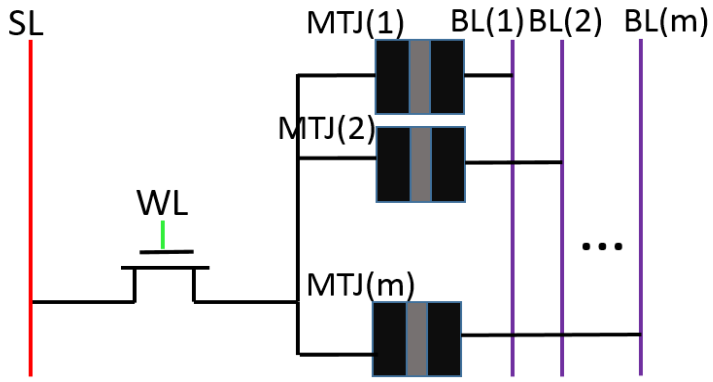


Figure 41 - Shared transistor cell structure for STT-MRAM [60].

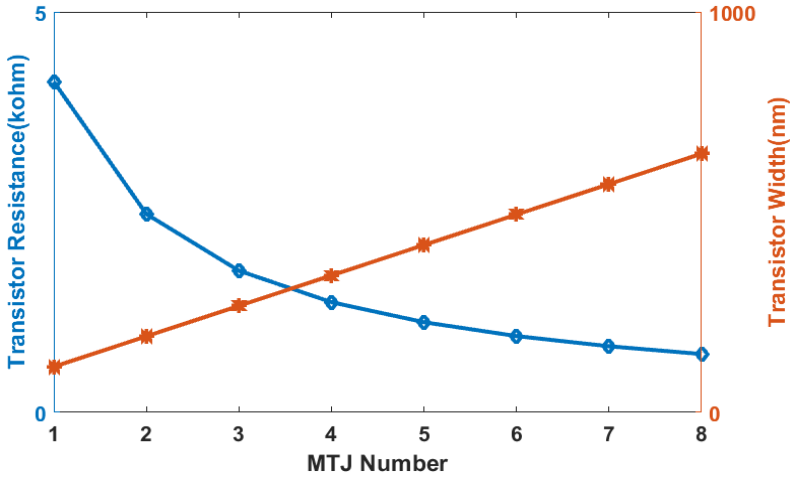


Figure 42 - Impact of multi-MTJ cell structure on cell transistor width and resistance in STT-MRAM.

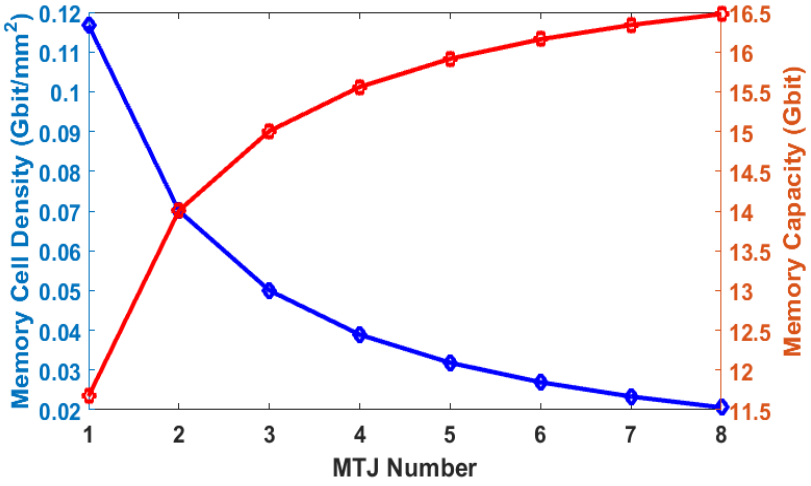


Figure 43 - Impact of multi-MTJ cell structure on the total number of memory cells and memory capacity assuming constant die area in STT-MRAM.

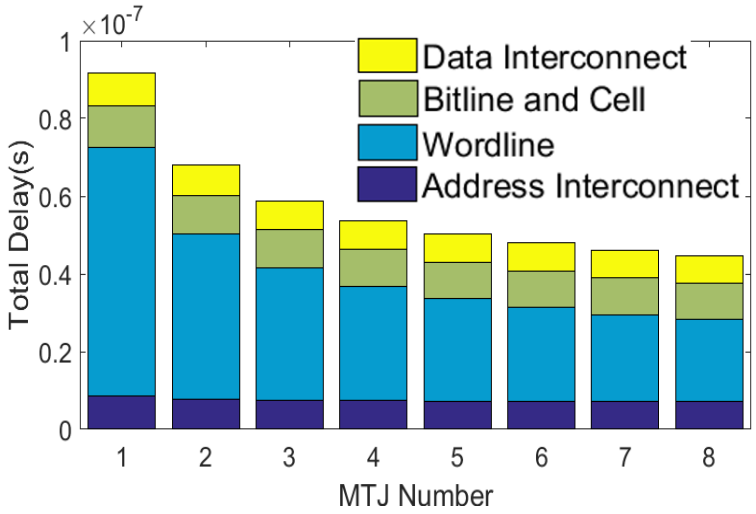


Figure 44 - Impact of multi-MTJ cell structure on the total memory latency assuming constant memory capacity in STT-MRAM.

CHAPTER 6 RESISTIVE RANDOM ACCESS MEMORY (ReRAM)

6.1 Introduction

As the CMOS technology advances to the deep nanoscale era, DRAM scaling faces serious challenges in speed, bandwidth, capacity, and cost [49]. NAND Flash technology has been the leading non-volatile memory technology for many years. However, it is believed that scaling this technology below 25nm has significantly degraded performance and reliability, thus, resulting in significant overhead complexity and computational power-demand from the system controller [53]. To find a solution, many emerging technologies, such as PRAM (Phase-Change RAM), MRAM (Magnetoresistive RAM), FeRAM (Ferroelectric RAM), and ReRAM (Resistive RAM), are being studied. Among these technologies, ReRAM is particularly promising. [61-65].

In addition to utilizing the memory resistor (memristor) technology to build the memory cell, RRAM arrays could be constructed using different arrangements such as the conventional 1T1R cell and the cross-bar cell arrangement. Section 6.3 and 6.4 present complete investigations of RRAM technology using the conventional 1T1R and cross-bar memory array structures. The benefits and disadvantages of each memory structure are studied and different ways to optimize the memory array are discussed.

6.2 Model Approaches and Assumptions

Table 5 shows the important parameters of the model.

Table 5 - Important parameters of the ReRAM model.

Parameter	Value	Ref.	Parameter	Value	Ref.
Chip size	10mm×10mm		Technology node	9.5nm	[24]
Cell architecture	1T1R		WL/BL/SL pitch	38/68/68 nm	
Cell size	38nm×136nm	[66]	M1 AR	2.0	[24]
FinFET(L/W)	20/135nm	[66]	M1 Resistance	122.907×10^{6} Ω /m	[24]
V_{dd} (core/IO)	0.8/1 V	[66]	M1 pitch	38nm	[24]
Memristor LR/HR	380Ω/ 360kΩ	[65]	F=M1 half-pitch	19nm	[24]
Memristor switching time	300ps	[65]			

6.2.1 Memory Cell Structure

In the conventional memory architecture, ReRAM cell consists of one access transistor and one memristor. The access transistor enables the connection between the sourceline (SL) connected to the cell memristor and the bitline (BL) which carries data to and from the memory cell. In this structure, the RRAM cell size is determined by the size of the large access transistor required for delivering the large write current to the memory cell. This prevents further increase in cell density and memory capacity.

6.2.2 Memory Subarray Architecture

The memory array structure and the architecture of global and local interconnects in ReRAM is similar to DRAM as explained in chapter 1, and shown in Figure 2 and Figure 3. The only difference is that the cell retaining element, which is a capacitor in DRAM, is a memristor in ReRAM. For writing a “0” value to a memory cell, a high

voltage is put on the WL, the BL is connected to V_{DD}, and the SL is connected to GND, which results in the cell memristor switching from the low-resistance (LR) state to the high-resistance state (HR). For writing a "1" to a memory cell, the BL and SL voltages are reversed, which causes the cell memristor to switch from HR to LR. The reading process is similar to writing a "0" value, but I_{Read} is much smaller than I_{Write} in order to avoid changing the resistance state of the memristor while reading the cell value.

There are different directions in the field of ReRAM research that include optimization of memristor, cell structure, memory configuration, layout, and interconnects. To find the most crucial research direction, the results for the memory performance are studied in order to find the bottlenecks of the ReRAM performance.

6.3 Model Results and Discussions

Figure 66 shows the total memory delay for the read and write operations. The delay of peripheral circuits such as decoders, multiplexers, and amplifiers have been included in the model, but their values are negligible and not visible in the figure. The main contributors to the memory delay are the local and global interconnects, and the cell. As for the interconnects delay, for a small number of banks, since the banks are large and the WLs and the BLs are long, most of the interconnect delay comes from the local interconnects inside the banks; i.e. the WL and BL. For a large number of banks, most of the interconnect delay comes from the global address and data interconnects. Increasing the number of banks reduces the local interconnects delay, but has little impact on the global interconnects delay.

Most of the dynamic power is consumed in the global address interconnects, data interconnects, and the cell as shown in Figure 67. Data interconnects dissipate far more energy than the address interconnects since they transmit decoded data, and as a result require a much larger number of wires. By increasing the number of memory banks, the total memory latency is reduced while the dynamic power consumption is increased as shown in Figure 66 and Figure 67. However, the increase in power consumption would not be problematic since, as seen in Figure , by increasing the number of banks, the access energy to a memory block is still reduced. This is due to the large reduction in the memory block access time. As the number of memory banks increases, the limitation comes from the added peripheral circuits area since each bank has its own set of peripheral circuits which include decoders, multiplexers, and sense amplifiers. As seen in Figure , we have only increased the number of banks to the point that the total peripheral circuits area does not exceed 10% of the memory chip area.

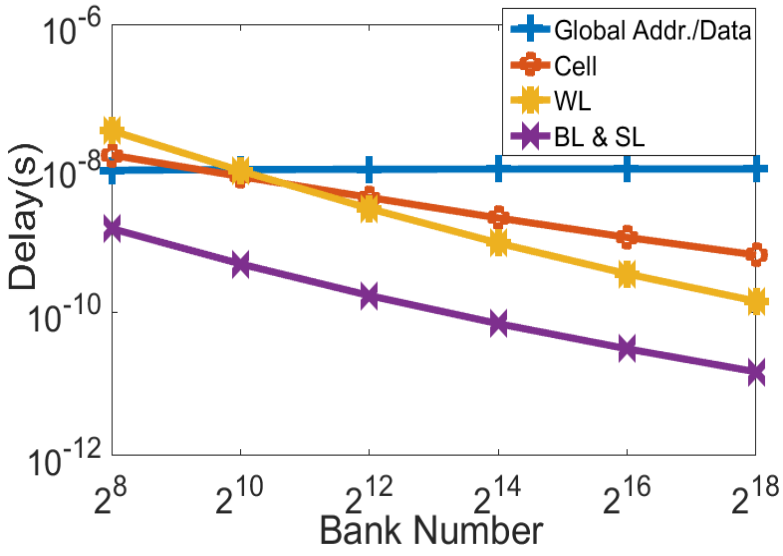


Figure 45 - Memory access time components in ReRAM array.

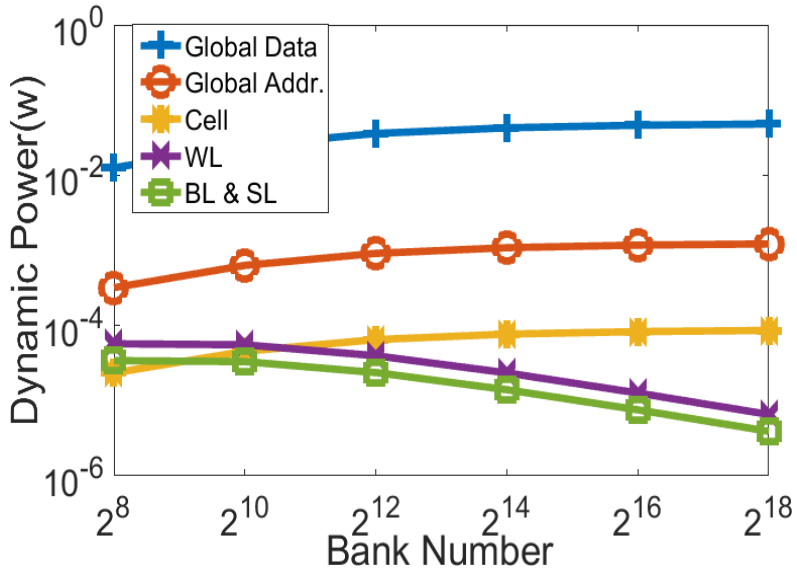


Figure 46 - Memory dynamic power consumption in ReRAM array.

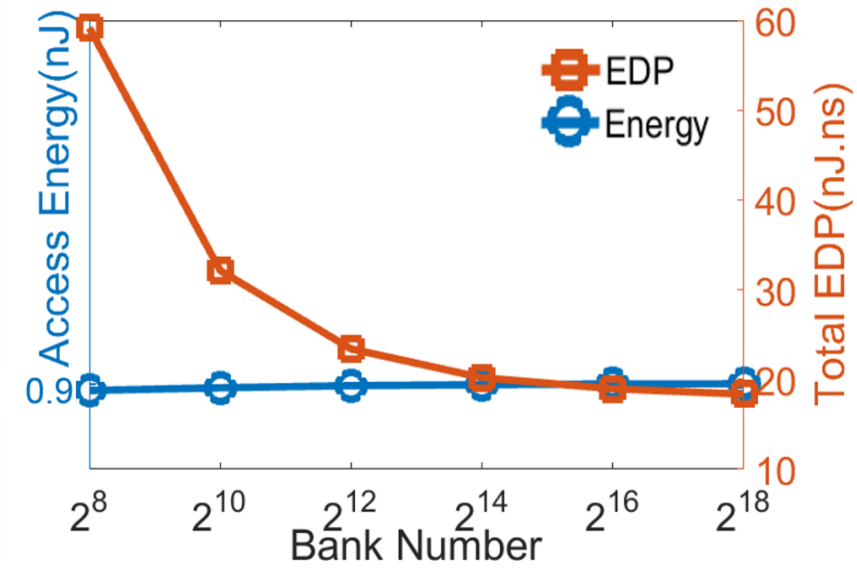


Figure 68 - Total memory access energy and EDP in ReRAM array.

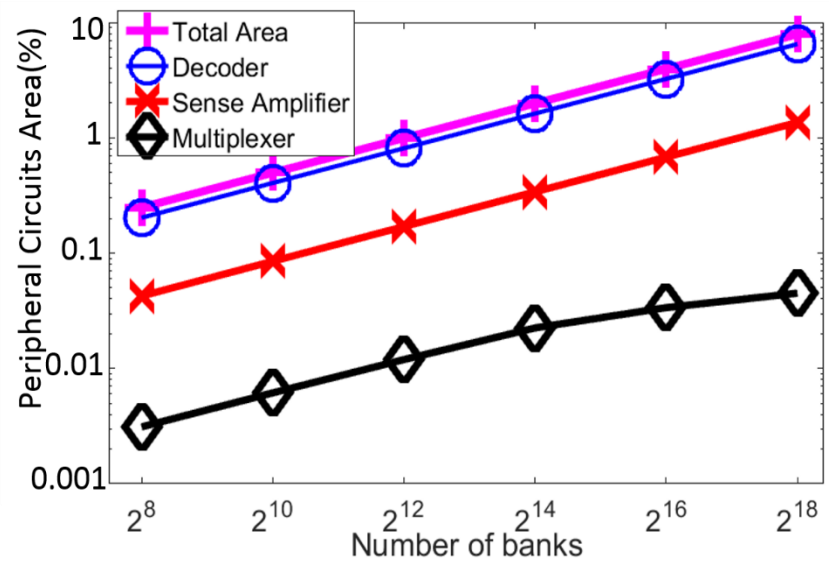


Figure 69 - Memory logic circuits area components in ReRAM array.

6.4 Cross-Bar ReRAM Array

As mentioned before, in the 1T1R memory cell structure, the cell size is determined by the cell access transistor size. The large size of the access transistor capable of delivering the large required current for switching the cell memristor prevents further scaling of the ReRAM cell and increasing the cell density. A solution to this problem is the cross-bar arrangement for memory array. In the cross-bar structure, the access transistors used to enable the connection between BL and cell memristor and the WLs connected to the access transistors are removed. The memory cells are placed at the intersections of the BLs and the SLs, and access to a memory cell is realized by applying different voltages to the BLs and the SLs.

The cross-bar memory structure has two important advantages. By reducing the cell size, it increases the cell density and memory capacity. In addition, by removing the WL and the access transistor, it reduces the memory latency and power consumption. The

relative improvement in memory delay offered by the cross-bar architecture depends on how dominant the WL delay is compared to 1T1R array. Figure 70 shows the delay components for the cross-bar ReRAM array. Figure 71 shows the total memory delays for ReRAM arrays using the conventional 1T1R and the cross-bar memory structures. As the number of banks increases, and banks get smaller, the WL delay becomes less dominant, and the difference between the delay of cross-bar and 1T1R arrays becomes smaller.

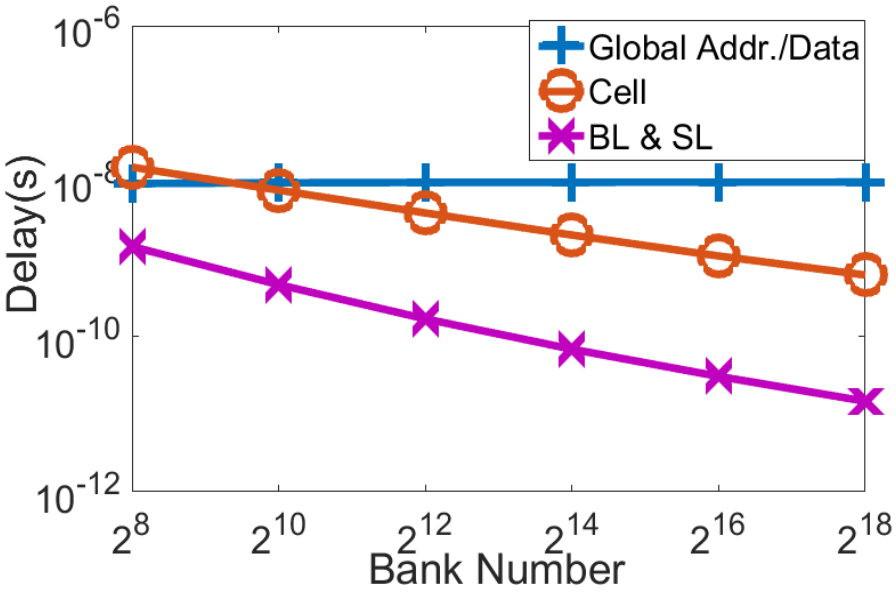


Figure 47 - Memory access time components for cross-bar ReRAM.

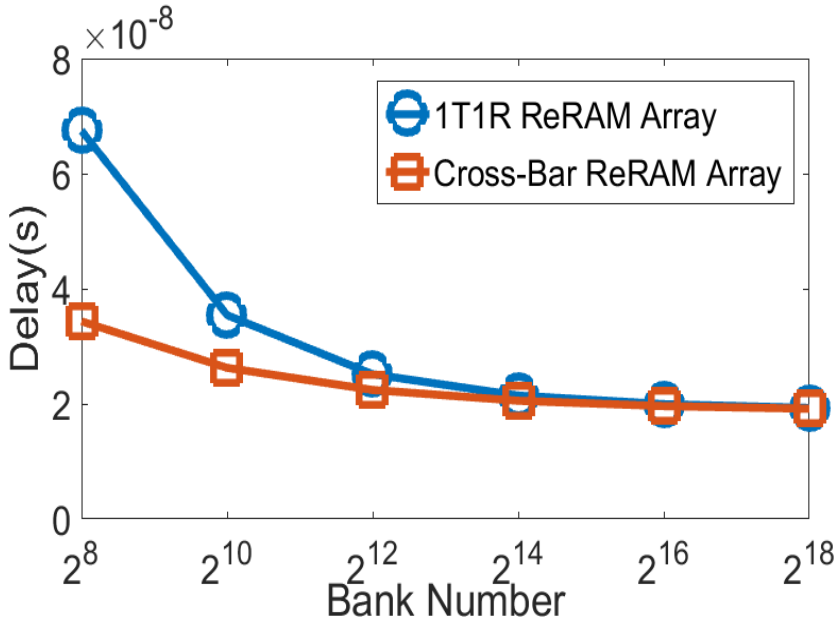


Figure 48 - Total memory latency for 1T1R and cross-bar ReRAM arrays.

As for the chip area, using the crossbar ReRAM technology, the cell size is only limited by local interconnects pitch inside the memory bank. This enables the ideal cell size of $4F^2$ for the memory cell where each cell side is only as wide as the local wire pitch (2F). At 9.5nm technology node, assuming die size of $100mm^2$, this increases the memory capacity by 250% from 2.0788 Gbit to 7.2758 Gbit.

In the cross-bar structure, for writing "1" to a cell, the SL connected to the cell is connected to GND and the BL is connected to V_{SET}. All the other SLs and BLs are connected to $V_{SET}/2$. This way, the memory cells are divided into three groups that have the voltages of zero, $V_{SET}/2$, and V_{SET} across them. The cells with $V_{SET}/2$ across them are a source of concern since they draw parasitic currents and reduce the available voltage across the selected cell memristor. To address this issue, devices with nonlinear I-V characteristics similar to diodes are integrated in series with the memristors that only

allow the passage of current when the voltage across the cell is $V_{SET,}$ and are called the cell selectors [67].

6.5 Memristor Characteristics

Even after eliminating the leakage current paths by using cell selectors, the voltage drop along the BL and SL could be problematic since it creates the need for high V_{SET} and V_{RESET} supply voltages. The severity of this issue depends on the ratio of the sum of the memristor resistance and the local interconnects resistance to the memristor resistance. As a result, this is especially challenging during the RESET operation when the memristor is at LR state. Figure 72 shows the required V_{SET} and V_{RESET} for the case of using memristor with LR and HR of 380Ω and 360 kΩ. As seen in Figure 72, even at the highest bank number of 2^{18}, the required V_{RESET} for a memristor with LR of 380 ohm is about 10V, which is too high for an on-chip voltage supply. To address this problem, either the banks should be made smaller with shorter BLs and SLs, or memristors with higher LR values should be utilized. In a 2D memory, the bank size is limited at two levels; i.e. design level and operation level. At the design level, as mentioned before and seen in Figure , the bank size reduction is limited by the added peripheral circuits area due to the increasing bank number, assuming a constant memory chip area. At the operation level, the bank size reduction is limited by the memory block size, and reducing the memory block size reduces the system throughput. Figure 73 shows the required V_{RESET} for different bank numbers and different memristor LR values. As the number of banks increases, and the banks get smaller, the BL and SL resistance becomes smaller, and the ratio of the sum of the memristor resistance and the BL and SL resistance to the memristor resistance approaches one. As a result, for large numbers of banks, the

difference of RESET voltage for different LR values becomes smaller. Figure 74 shows the minimum limit for memristor LR in order to keep V_{RESET} below 3V.

One of the areas that the cross-bar ReRAM could stand out from the previous memory technologies is that in this structure, by adding metal levels, the BL and SL pair levels are built on top of each other with memory cells forming at the interconnects junctions. With this technology, a 3D cross-bar ReRAM array is built, and the bank footprint area could be reduced. As a result, V_{RESET} could be reduced without reducing the bank capacity or the memory block size.

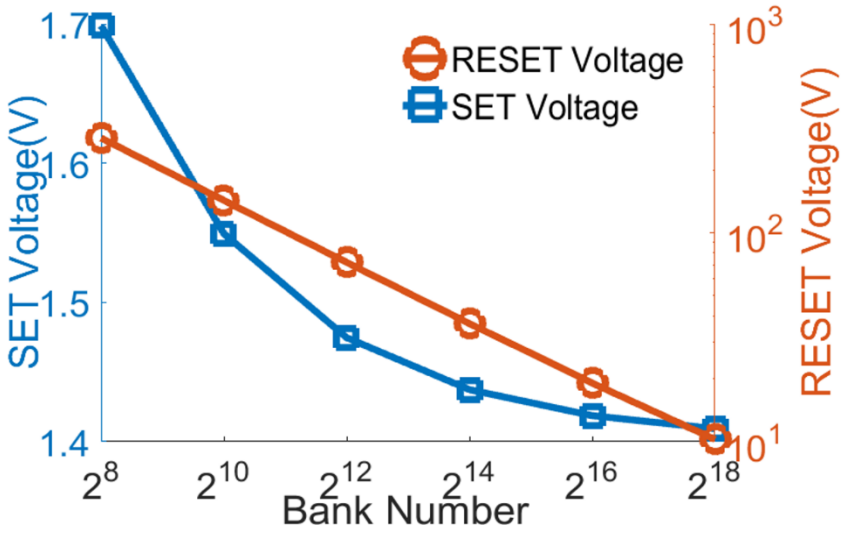


Figure 49 - Required SET and RESET voltages for cross-bar ReRAM array with different bank numbers and memristor LR=380Ω and HR=360kΩ.

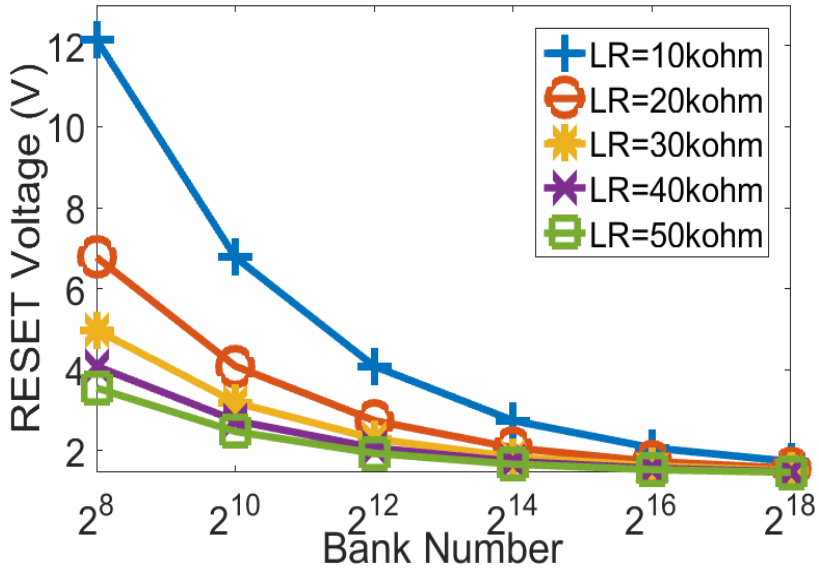


Figure 50 – Required RESET voltages for cross-bar ReRAM array with different bank numbers and memristor HR=360kΩ and different LR values.

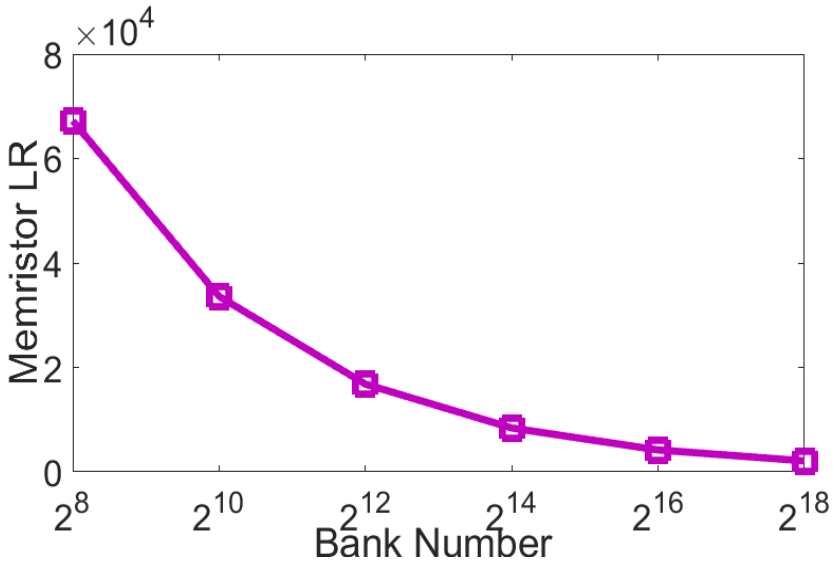


Figure 51 - Minimum limit for memristor LR to keep the required V_{RESET} below 3V in cross-bar ReRAM array.

The advantages of using a memristor with higher LR are lower required V_{RESET}, lower dynamic power consumption, and lower area of chip peripheral circuits since the memory array could be divided into a smaller number of larger banks. The disadvantages

of increasing the memristor LR is an increase in the cell delay. Figure 75 and Figure 76 show the impact of memristor LR on memory cell delay and power consumption. Increasing the memristor LR has the largest impact on the cell delay and power consumption during the cell read operation when the cell memristor is in LR state, and the memory cell has the "1" value. The cell write delay and power consumption are not affected considerably by the LR value since for the write operation, the memristor resistance changes from LR to HR or vice versa, and it is the worst case HR value that determines the overall delay because multiple bits of ones and zeros are written simultaneously.

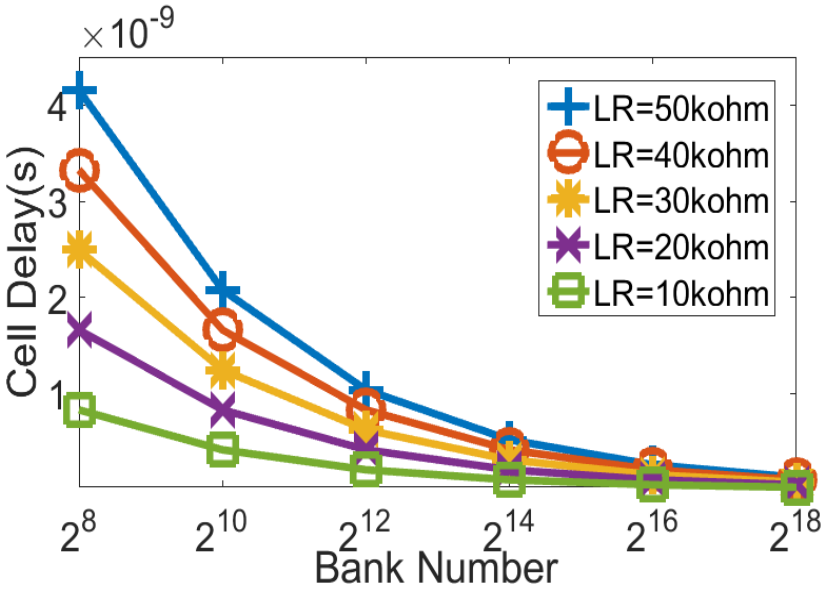


Figure 52 - Cell delay during the read operation of memory cell with "1" value for different values of memristor LR and memory bank number.

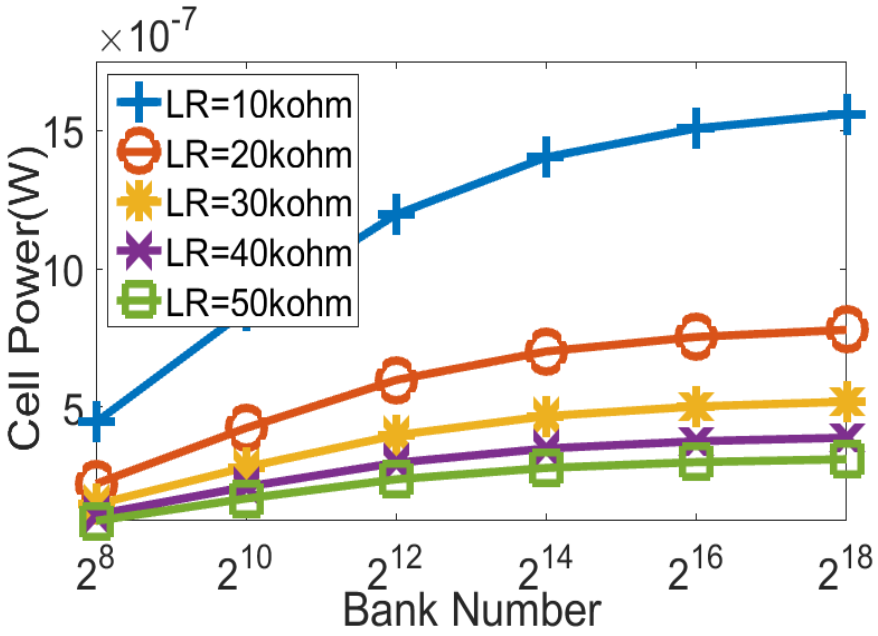


Figure 53 - Cell dynamic power consumption during the read operation of memory cell with "1" value for different values of memristor LR and memory bank number.

CHAPTER 7. GRAPHENE NANORIBBON (GNR) INTERCONNECTS IN MEMORY ARRAYS

7.1 Introduction

To alleviate the ever increasing performance gap between devices and interconnects, the graphene interconnect is one of the candidates that can potentially outperform the conventional copper wires thanks to its outstanding electrical properties including the long electron mean free path (MFP), the large current conduction capacity, and the small capacitance per unit length. However, since graphene is a two-dimensional structure, increasing the interconnect pitch does not lower resistance as fast as it does for copper wires. Hence, comparing graphene and copper interconnects strongly depends on the interconnect pitch. On the other hand, a multilevel interconnects accommodates interconnects of various lengths routed in metal levels with very different wiring pitches. As a result, to better understand and evaluate the overall benefits of using graphene interconnects, circuit- and system-level analyses are essential. The studies are done for DRAM and STT-MRAM memory technologies.

7.2 Model Results and Discussions

Table 6 shows the important parameters of copper and graphene interconnects used in our model.

Table 6 - Important parameters of the memory model and copper and graphene interconnects.

Paremeter	Value	Parameter	Value
Chip size (mm^2)	100	Graphene Interconnects	
Technology node (nm)	9.5	E_f(eV)	0.4
DRAM cell size	$16F^2$	Width(nm)	20
STT-MRAM cell size	$40F^2$	Edge scattering	0.2
Copper interconnects		Temperature(K)	300
Copper wire pitch(nm)	38	Mean free path(nm)	300
Local Interconnects AR	2.0	Number of layers	10
Global Interconnects AR	2.34	Contact R(Ω×μm)	100

Figure 77 and Figure show the memory latency components of a DRAM memory array using copper and graphene interconnects, respectively.

For the global copper interconnects, optimal repeaters are found to be the around 2000 and minimum size inverters. The number of repeaters used for global grapehene interconnects is 2-3 times the number used for copper interconnects. These interconnects are about 15mm long. For global interconnects, if 2-3 times the number of repeaters used for copper interconnects are used for grapehen interconnects, the graphene interconnect delay is about twice of the copper interconnect delay. The width is about the same, Copper wire half-pitch is 19nm, and GNR width is 20nm. Copper interconnect pitch was chosen as 36nm instead of 40nm in order to correspond to the ITRS 9.5nm technology node.

For both copper and graphene local interconnects, no repeaters are used due to the cell size limit. Elmore delay model is used, and the input capacitance of the transistors driven by the wordline is added to the wordline capacitance. As seen in the Figure 77 and Figure , the graphene wordline delay is 14 times the copper wordline delay, with the same width for copper wire and GNR.

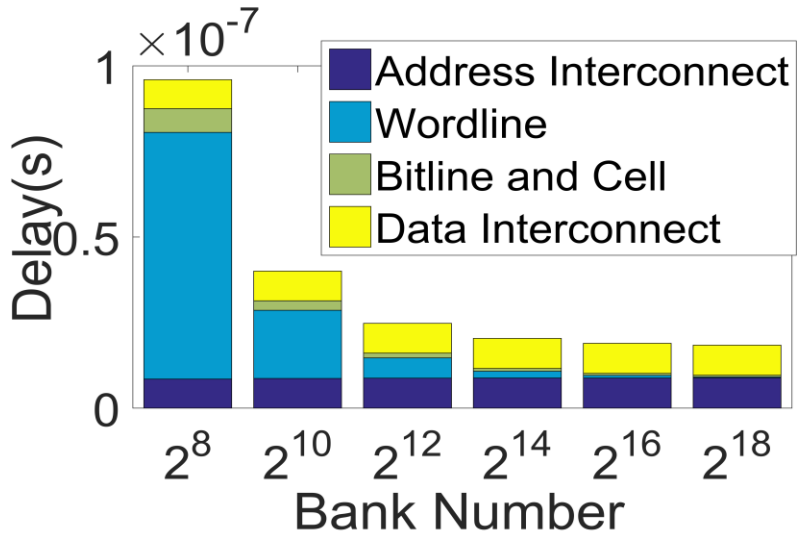


Figure 54 - Memory latency components of DRAM using copper for all the local and global interconnects.

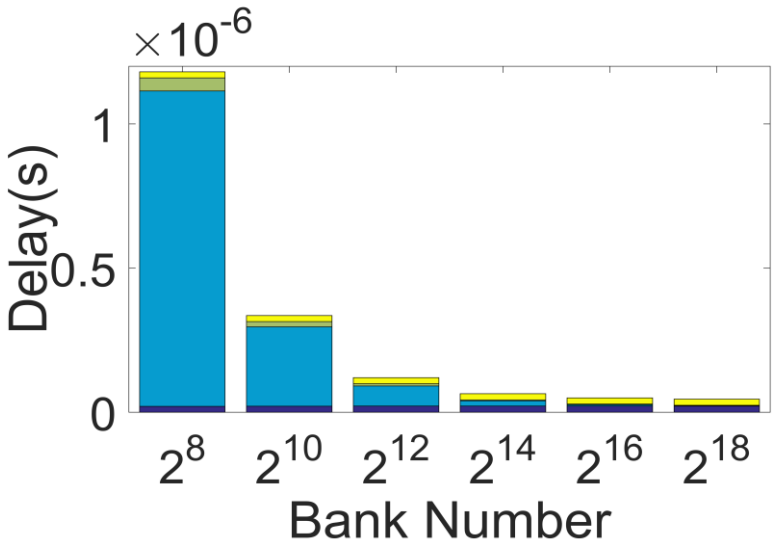


Figure 78 - Memory latency components of DRAM using graphene for all the local and global interconnects.

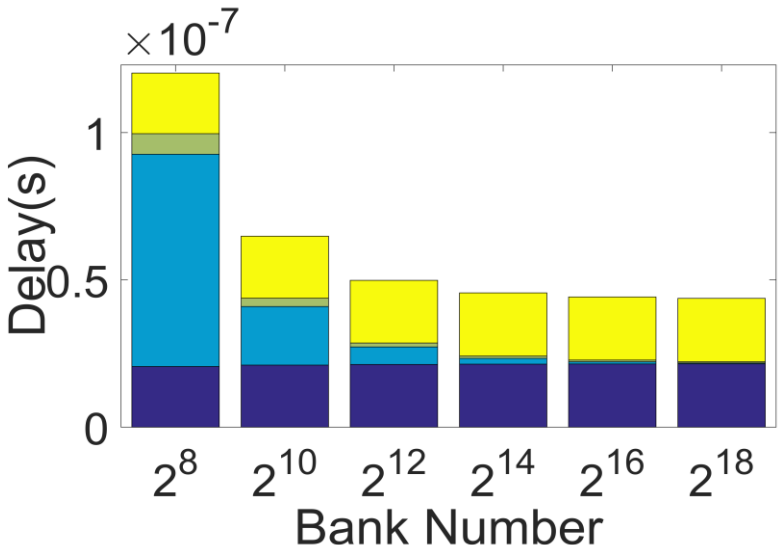


Figure 79 - Memory latency components of DRAM using graphene for global address and data interconnects, and copper for local wordlines and bitlines.

In conclusion for DRAM memory arrays: 1) Replacing copper with graphene in the global address and data interconnects can potentially reduce the memory latency using optimal values for graphene fabrication parameters including width, number of layers, and electron MFP. 2) Using graphene for local wordlines and bitlines increases the delay

by up to a couple of orders of magnitude, and should be avoided. This increase in delay is due to the higher sheet resistance and contact resistance of graphene.

Using these results, an optimum interconnect structure for reducing the DRAM memory latency is presented. In this structure, graphene is used for the global address and data interconnects, and copper is used for the local wordlines and bitlines. Using graphene for the local intreconnects increases the delay significantly due to graphene's higher resistance compared to copper. The delay of this optimum interconnect structure for DRAM is shown in Figure .

Figure 80 and Figure 81 show the DRAM dynamic power using copper and graphene interconnects. Figure 82 compares these results with the result of using the optimal hybrid interconnect structure using graphene for global interconnects and copper for local interconnects. As seen in these figures, most of the dynamic power is consumed in the global data interconnects due to the large number of wires. As a result, the total power only depends on the material used for the global interconnects. Using graphene for global address and data inerconnects reduces the memory power by up to 65%. The plots of dynamic power for the DRAM and STT-MRAM arrays are similar, since the interconnects are quite similar for different memory technologies, and the main difference is the cell structure. As a result, for STT-MRAM, the main contributor to the memory power are the global data interconnects also. However, compared to DRAM, the bitline and cell power is larger in STT-MRAM arrays.

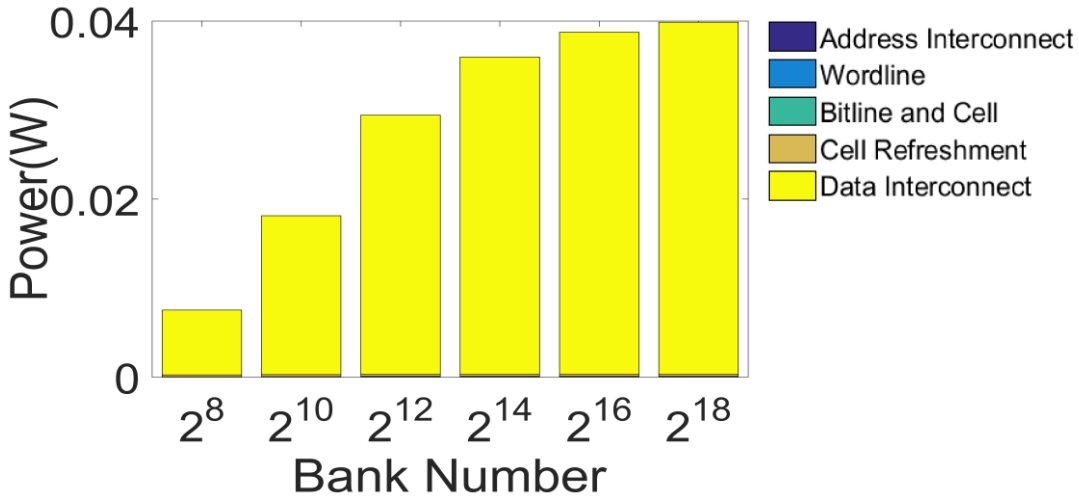


Figure 55 - DRAM dynamic power dissipation components using copper for all the local and global interconnects.

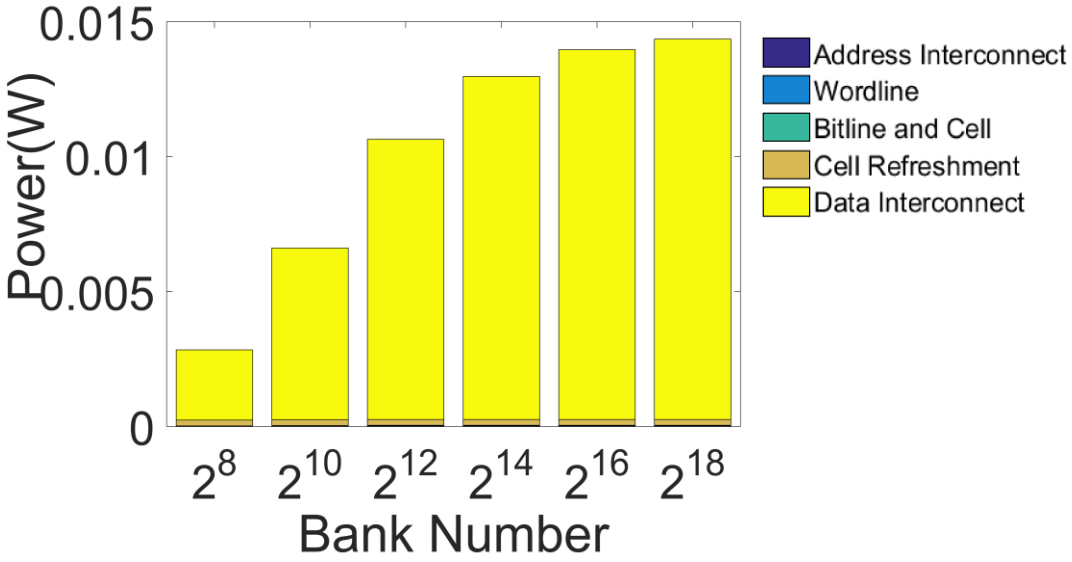


Figure 56 - DRAM dynamic power dissipation components using graphene for all the local and global interconnects.

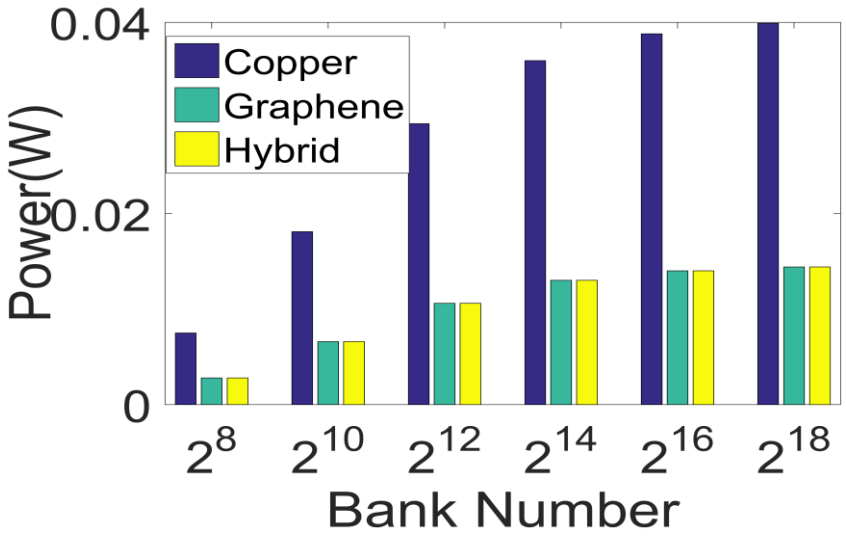


Figure 57 - DRAM total dynamic power dissipation using copper, graphene, and hybrid interconnect structures.

7.3 Impact of GNR Parameters on Memory Latency

In this section, we study the impact of GNR parameters such as graphene width, number of layers, and MFP on the DRAM array latency. The studies are done for the memory array with the interconnect structure using graphene for the global address and data interconnects and copper for the local wordlines and bitlines. As shown in Figure 83, increasing the graphene width from 20nm to 200nm reduces the global interconnects delay by 86% and the total memory latency by 47%. Figure 84 shows that there is an optimum number of graphene layers for minimzing the interconnect delay. With the parameter values used in our model, the optimum number of graphene layers is 40 which reduces the global interconnects delay by 91% and the total memory latency by 67%. If the electron MFP is increased by 10 times from 100nm to 1000nm, the results would be 45% reduction in global interconnects delay and 17% reduction in total memory latency as shown in Figure 85. Finally, Figure 86 shows the impact of graphene interconnect electron edge scattering on the global interconnect delay and total memory latency.

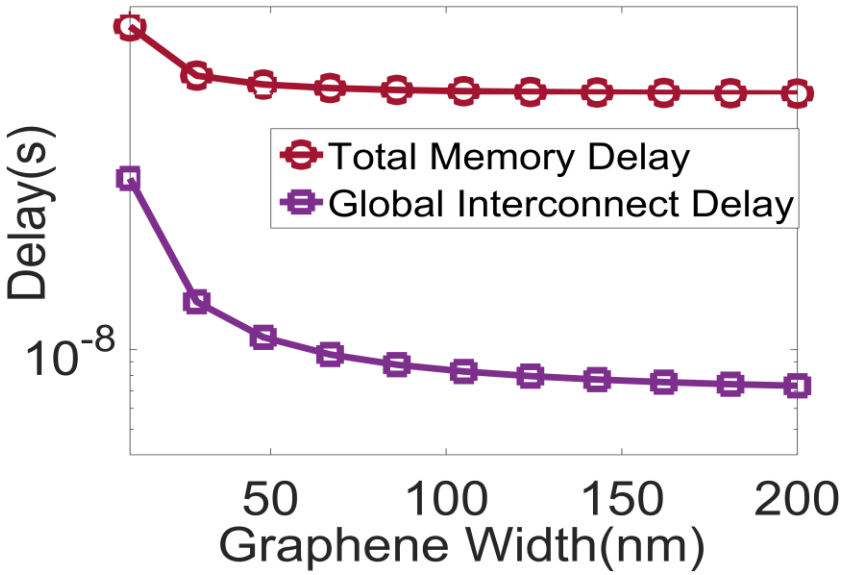


Figure 58 - Impact of global graphene interconnects width on the global interconnects delay and total memory latency.

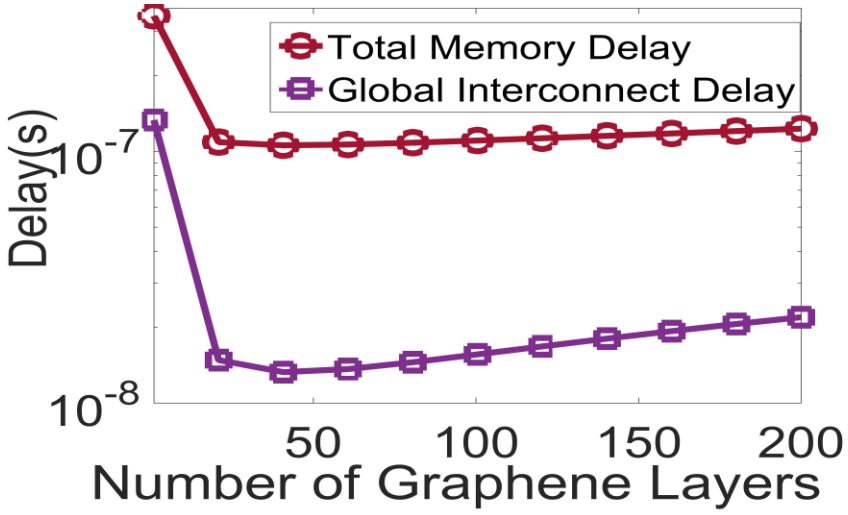


Figure 59 - Impact of global graphene interconnects number of layers on the global interconnects delay and total memory latency.

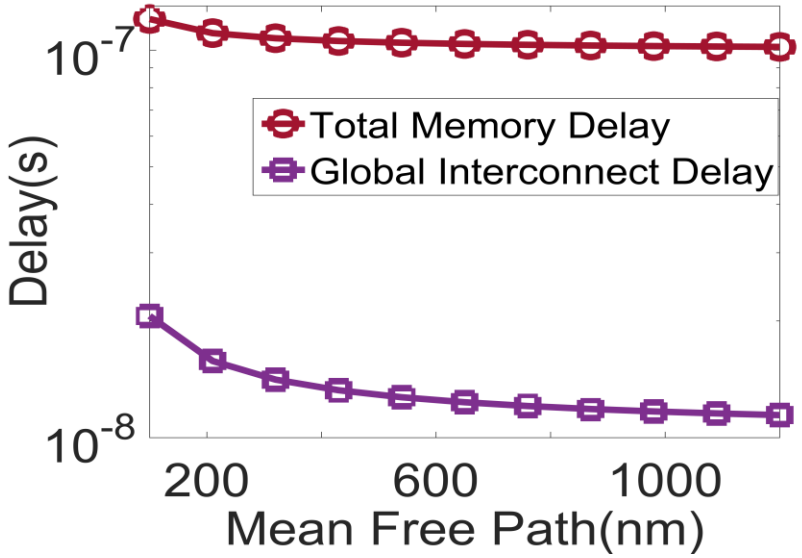


Figure 60 - Impact of global graphene interconnects electron MFP on the global interconnects delay and total memory latency.

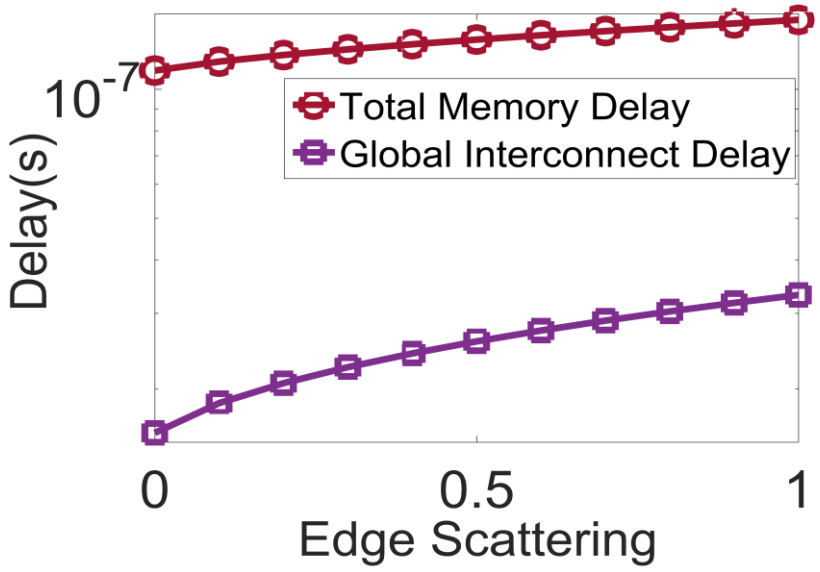


Figure 61 - Impact of global graphene interconnects electron edge scattering on the global interconnects delay and total memory latency.

CHAPTER 8 CONCLUSION AND FUTURE WORK

In the past, only the long global interconnects imposed limits on the chip clock frequency since the delay of local interconnects scaled with technology. However, the increase in the copper resistivity due to size effects [44] such as surface/grain boundary scattering, and line edge roughness have led to a significant increase in local interconnect delay causing it to become an important limiting factor, too [45]. Interconnects dissipate more than 50% of the dynamic power in a microprocessor [29]. To reduce the interconnect capacitance, the industry has made major investments in low-k dielectrics to mitigate the impact of the ever increasing total interconnect length. However, due to mechanical and reliability issues, the scaling of the effective dielectric constant has virtually stopped at around 2.5 to 2.7. It is believed that it may not be possible to achieve dielectric constants below 2 [24].

In multi-core systems, the memory latency and bandwidth are among the key limitations. Therefore, modeling and benchmarking the interconnect performance for memory chips is of utmost importance. The memory system design is facing many challenges. DRAM-based memory systems are stretched to meet the increasing demands on power efficiency, high memory bandwidth and large memory capacity that are required by multi-core processors. To address these challenges both technology and circuit solutions are investigated. Chapter 1 presented a more comprehensive introduction and background.

In chapter 2, the structure of the memory system used in the models was broken down into its constituents and explained in detail. Different kinds of interconnects in the memory array were introduced. The chip layout and all the metal levels were

demonstrated. Different decoding methods were investigated and compared comprehensively. Memory logic circuits were studied, modeled, and parametrical models were presented for the footprint area.

In chapter 3, the impact of interconnects on the performance and power dissipation of DRAM array at various generations of technology was quantified. Various design aspects, such as the number and aspect ratio of banks, decoder driving capability, and the number of wiring levels, were optimized to minimize the overall EDP across various technology generations. It was found that at the ITRS 9.5nm technology node, adding two metal levels can decrease the memory delay by 18%, changing the banks aspect ratio from 1:1 to 2:1 reduces the delay by 40%, increasing the decoder size from minimum size to twice of that results in 37% delay reduction, and developing a single crystal copper interconnect technology leads to 21% delay reduction. Using Al or hybrid Cu-Al interconnect technologies were also studied, which lead to up to 28% reduction in the memory access time. Finally, the scaling analyses showed insightful trends of the bottlenecks of delay and power dissipation of a memory system down to the 3nm technology node.

To address the challenges associated with DRAMs performance, cost and scaling, both technology and circuit solutions should be investigated. Chapter 4 was dedicated to the investigation of one of the most promising solutions to these challenges which is 3D memory integration.

In chapter 4, three 3D memory configurations were investigated which reduce the memory delay and footprint area by up to 37% and 75%, respectively, in a memory chip with four stacked dies. 3D integration at memory array level and memory bank level was

studied. An optimized 3D memory configuration was designed which reduces the memory access time by 33% while maintaining small added area due to TSVs. The impact of interconnects on the performance and power dissipation of the 3D DRAM array at various generations of technology was quantified. The solutions to memory array scaling enabled by 3D integration were investigated. It was demonstrated that by die-stacking and reducing the bitline length and capacitance, the cell storage capacitance could be reduced down to 5fF which reduces the total memory delay by 45%. The impacts of the TSV and MIV, used in die-stacked 3D ICs and monolithic 3D ICs respectively, on the memory array performance and area along with their fabrication challenges were investigated. Reducing TSV diameter from 20μm to 2μm reduces the TSVs area by 99% and the area of a memory chip with four stacked dies by 42%. However, the challenges of the TSV fabrication at small diameters including the difficulty of aligning the stacked dies and the vias should be considered. It was shown that at 9.5nm technology node, the MIV delay is five orders of magnitude smaller than the memory access time which is around 10^{-8}s. As a result, if the MIV fabrication challenges are overcome, the MIV diameter could be reduced to 100nm without deteriorating the memory access time and frequency. The scaling trends of memory systems with different configurations were studied down to the ITRS 3nm technology node. The benefits of the transition from the 2D to the 3D memory for different memory array architectures were also quantified.

Emerging non-volatile memory technologies are being investigated as potential solutions, and STT-MRAM is one of the promising technologies among them. Remarkable progresses in STT switching with MgO as the magnetic tunnel junction (MTJ) barrier and increasing interest in STT-MRAM in the semiconductor industry have been witnessed in recent years. In chapter 5, a comprehensive interconnect analysis was presented for STT-MRAM chips, and the limits interconnects impose on the total memory delay, area, and dynamic power consumption were quantified. It was found that global and local interconnects constitute up to 80% of the memory delay. The global interconnects delay can be reduced by increasing the wiring pitch and the number of metal levels. Increasing the number of metal levels by 4 results in 28% reduction in memory latency. For the local interconnects, different techniques including increasing the decoder drive current, and changing the memory bank aspect ratio could reduce the total delay by 22% and 20% respectively. STT-MRAM reliability challenges including interconnects lifetime reduction by EM were also studied. Finally, the impacts of a potential alternative multi-bit cell structure on the memory performance was investigated, which include up to 53% reduction in memory latency for a constant capacity, or 44% increase in memory capacity for a constant die area.

NAND Flash technology has been the leading non-volatile memory technology for many years. However, it is believed that scaling this technology below 25nm has great challenges. Therefore, the search for a novel non-volatile memory technology is crucial. Among the many emerging non-volatile memory technologies, Resistive RAM (ReRAM) is particularly promising. In chapter 6, a comprehensive interconnect analysis was presented for ReRAM chips, and the limits interconnects impose on the total memory

delay, area, and dynamic power consumption were quantified. It was found that global and local interconnects constitute up to 80% of the memory delay. Two ReRAM technologies of 1T1R and cross-bar array were compared. The advantages and challenges of cross-bar ReRAM arrays were investigated, and potential solutions were presented. Impacts of the memristor characteristics on the ReRAM performance were studied. Other challenges of the cross-bar structure such as lower read margin and higher leakage power resulting from the non-ideality of cell selectors were among the topics in ReRAM design that can be investigated further.

Due to the ever increasing copper resistivity at ultra-small technology nodes, the search for novel materials for interconnects is of utmost importance. The graphene interconnect is one of the candidates that can potentially outperform the conventional copper wires thanks to its outstanding electrical properties including the long electron mean free path (MFP), the large current conduction capacity, and the small capacitance per unit length. Chapter 7 studied the use of the graphene interconnects as local and global interconnects in DRAM and STT-MRAM technologies. It was found that replacing copper with graphene in the global address and data interconnects can potentially reduce the memory latency using optimal values for graphene fabrication parameters including width, number of layers, and electron MFP. However, using graphene for local wordlines and bitlines increases the delay by up to a couple of orders of magnitude, and should be avoided. This increase in delay is due to the higher sheet resistance and contact resistance of graphene.

Using these results, an optimal interconnect structure for reducing the memory latency in DRAM and STT-MRAM was presented which includes using graphene for the

global address and data interconnects and copper for local wordlines and bitlines. This optimal structure reduces the DRAM array latency by up to 60%. In addition, the impact of using graphene interconnects on the memory array power dissipation was investigated. Due to the large power consumption of the global data interconnects, the total memory power dissipation depends only on the material used for global interconencts. Using graphene for global address and data inerconnects reduces the memory power by up to 65%. The plots of dynamic power for the DRAM and STT-MRAM arrays are similar, since the interconnects are quite similar for different memory technologies, and the main difference is the cell structure. As a result, for STT-MRAM, the main contributor to the memory power are the global data interconnects also. However, compared to DRAM, the bitline and cell power is larger in STT-MRAM arrays. Finally, the impact of GNR parameters such as graphene width, number of layers, and MFP on the DRAM array latency was investigated.

The continuation of the research in the interconnect optimization for memory systems should be done by adding to both the depth and breadth of the work. For the memory technologies and interconnect structures investigated in this work, there are numerous topics that lack of time didn't allow us to cover in this work. Some of the current exciting research topics include transistor and capacitor leakage current reduction techniques, low latency and energy-efficient DRAM cache designs, hybrid DRAM-NVM memory architecture, power management in 3D DRAM-stacked memory, efficient fault tolerance methods, data privacy in non-volatile memories, impact of process variation, reliability challenges, impact of non-ideal characteristic of cell selectors in cross-bar memory arrays, novel materials for MTJs in STT-MRAM arrays, etc.

In addition to the the memory and interconnect technologies covered in this work, some of the emerging device and interconnect technologies pursued in order to extend Moore's law to beyond-2020 technology generations include carbon–based devices [15-16] and interconnects [17-18], nano–electromechanical systems (NEMS) [19], optical or photonic interconnects [20-21], and even non–charge–based systems [22]. Also some of the novel memory technologies include PRAM (Phase-Change RAM), MRAM (Magnetoresistive RAM), and FeRAM (Ferroelectric RAM).

REFERENCES

[1] G. E. Moore, “Cramming More Components onto Integrated Circuits”, Reprinted from Electronics, volume 38, number 8, pp.114-116, April 19, 1965.

[2] D. Brock, “Understanding Moore’s Law: Four Decades of Innovation”, Chemical Heritage Foundation, pp. 67–84, ISBN 0-941901-41-6, 2006.

[3] M. Kanellos, "Moore's Law to roll on for another decade”, http://news.cnet.com/2100-1001-984051.html.

[4] D. Buchanan, “Scaling the gate dielectric: materials, integration and reliability,” IBM Journal of Research and Development, vol. 43, pp. 245–264, May 1999.

[5] Y. Yeo, Q. Lu, W. Lee, T.-J. King, C. Hu, X. Wang, and T. Ma, “Direct tunneling gate leakage current in transistors with ultrathin silicon nitride gate dielectric,” IEEE Electron Device Letters, vol. 21, pp. 540–542, November 2000.

[6] P. Bai et al., “A 65nm logic technology featuring 35nm gate lengths, enhanced channel strain, 8 Cu interconnect layers, low-*k* ILD and 0.57 *m*m2 SRAM cell”, IEDM Technical Digest, pp. 657–660, December 2004.

[7] D. Antoniadis, I. Aberg, C. Ni Chleirigh, O. Nayfeh, A. Khakifirooz, and J. Hoyt, “Continuous MOSFET performance increase with device scaling: the role of strain and channel material innovations,” IBM Journal of Research and Development, vol. 50, pp. 363–376, July/September 2006.

[8] R. Chau, S. Datta, M. Doczy, J. Kavalieros, and M. Metz, “Gate dielectric scaling for high–performance CMOS: from SiO2 to high–*k*,” in Intl. Workshop on Gate Insulator, pp. 124–126, November 2003.

[9] G. D. Wilk, R. M. Wallaceb, J. M. Anthony, “High-k gate dielectrics: Current status and materials properties considerations”, Journal of Applied Physics, Vol. 89, pp. 5243-5275, 15 May 2001.

[10] C. Auth et al., “45nm high-–*k* + metal gate strain–enhanced transistors,” in Symposium on VLSI Technology, pp. 128–129, June 2008.

[11] C. Auth et al., “A 22nm high–performance and low–power CMOS technology featuring fully–depleted tri–gate transistors, self–aligned contacts and high density MIM capacitors,” in Symposium on VLSI Technology, pp. 131–132, June 2012.

[12] J. Meindl et al., “Interconnecting device opportunities for gigascale integration (GSI),” in IEDM Technical Digest, pp. 23.1.1–23.1.4, December 2001.

[13] D. Edelstein et al., “Full copper wiring in a sub–0.25 *m*m CMOS ULSI technology,” in IEDM Technical Digest, pp. 773–776, December 1997.

[14] M. Bohr, “The new era of scaling in an SoC world,” in IEEE International Solid State Circuits Conference, pp. 23–28, February 2009.

[15] Y.-M. Lin et al., “100–GHz Transistors fromWafer Scale Epitaxial Graphene,” Science, vol. 327, p. 662, February 2010.

[16] A. Bachtold, P. Hadley, T. Nakanishi, and C. Dekker, “Logic circuits with carbon nanotube transistors,” Science, vol. 294, pp. 1317–1320, October 2001.

[17] A. Naeemi and J. Meindl, “Design and performance modeling for single–walled carbon nanotubes as local, semiglobal, and global interconnects in gigascale integrated systems,” IEEE Transactions on Electron Devices, vol. 54, pp. 26–37, January 2007.

[18] A. Naeemi and J. Meindl, “Compact physics–based circuit models for graphene nanoribbon interconnects,” IEEE Transactions on Electron Devices, vol. 56, pp. 1822–1833, September 2009.

[19] R. Nathanael, V. Pott, H. Kam, J. Jeon, and T.-J. Liu, “4–terminal relay technology for complementary logic,” in IEDM Technical Digest, pp. 1–4, December 2009.

[20] R. Beausoleil et al., “Nanoelectronic and nanophotonic interconnect,” in Proceedings of the IEEE, vol. 96, pp. 230–246, February 2008.

[21] A. Krishmamoorthy et al., “Computer systems based on silicon photonic interconnects,” in Proceedings of the IEEE, vol. 97, pp. 1337–1361, July 2009.

[22] B. Behin-Aein, D. Datta, S. Salahuddin, and S. Datta, “Proposal for an all–spin logic device with built–in memory,” Nature Nanotechnology, pp. 266–270, February 2010.

[23] H. Bakoglu and J. Meindl, “Optimal interconnection networks for ulsi,” IEEE Transactions on Electron Devices, vol. 32, pp. 903–909, May 1985.

[24] International Technology Roadmap for Semiconductors, 2017.

[25] D. Miller, “Rationale and challenges for optical interconnects to electronic chips,” Proceedings of the IEEE, vol. 88, pp. 728–749, June 2000.

[26] D. Miller, “Device requirements for optical interconnects to silicon chips,” Proceedings of the IEEE, vol. 97, pp. 1166–1185, July 2009.

[27] K. Koo, H. Cho, P. Kapur, and K. Saraswat, “Performance comparisons between carbon nanotubes, optical, and Cu for future high performance on–chip interconnect

applications," IEEE Transactions on Electron Devices, vol. 54, pp. 3206–3215, December 2007.

[28] C. Moore, "Data processing in exascale–class computer systems," in The Salishan Conference on High Speed Computing, April 2011.

[29] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect power dissipation in a microprocessor," in International Workshop on System Level Interconnect Prediction, pp. 7–13, 2004.

[30] A. Naeemi, A. Ceyhan, V. Kumar, C. Pan, R. M. Iraei, and S. Rakheja, "BEOL scaling limits and next generation technology prospects," in Proceedings of the 51st Annual Design Automation Conference, 2014, pp. 1-6..

[31] C. Pan and A. Naeemi, "A Paradigm Shift in Local Interconnect Technology Design in the Era of Nanoscale Multigate and Gate-All-Around Devices," Electron Device Letters, IEEE, vol. 36, pp. 274-276, 2015.

[32] Klaus Schuegraf, Mathew C. Abraham, Adam Brand, Mehul Naik, and Randhir Thakur, "Semiconductor logic technology innovation to achieve sub-10 nm manufacturing", IEEE Journal of the Electron Devices Society, Vol. 1, No. 3, March 2013.

[33] C. Pan and A. Naeemi, "A Proposal for a Novel Hybrid Interconnect Technology for the End of Roadmap," Electron Device Letters, IEEE, vol. 35, pp. 250-252, 2014.

[34] P. Garrou, "Handbook of 3D Integration", John Wiley & Sons, 2008.

[35] C. Pan and A. Naeemi, "System-level analysis for 3D interconnection networks," Proceedings of the IEEE Interconnect Technology Conference (IITC), 2013, pp. 1-3.

[36] M. Taouil, S. Hamidioui, J. Verbree, E. J. Marinissen, "On Maximizing the Compound Yield for 3D Wafer-to-Wafer Stacked ICs", Proceedings of the IEEE International Test Conference (ITC), 2010, pp. 1-10.

[37] D. Milojevic, P. Marchal, E. J. Marinissen, G. Van der Plas, D. Verkest, E. Beyne, "Design Issues in Heterogeneous 3D/2.5D Integration", Proceedings of the IEEE International Asia and South Pacific Design Automation Conference (ASP-DAC), 2013, pp. 403-410.

[38] J. A. Davis, et al., "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century", Proceedings of the IEEE, Vol. 89, Issue 3, pp. 305-324, 2001.

[39] K. H. Koo, P. Kapur, and K. C. Saraswat, "Compact Performance Models and Comparisons for Gigascale On-Chip Global Interconnect Technologies", IEEE Transactions on Electron Devices, Vol. 56, Issue 9, pp. 1787-1798, 2009.

[40] K. Banerjee, S. J. Souri, and K. C. Saraswat, "3-D ICs: A Novel Chip design for Improving Deep-Submicrometer Interconnect Performance and Systems-On-Chip Integration", Proceedings of the IEEE, Vol. 89, Issue 5, pp. 602-633, 2001.

[41] J. H. Lau, "Evolution, Challenge, and Outlook of TSV, 3D IC Integration and 3D Silicon Integration", Proceedings of the IEEE International Symposium on Advanced Packaging Materials, pp. 462-488, 2011.

[42] M. Taouil, M. Mosadeh, S. Hamdioui, and E. J. Marinissen, "Interconnect Test for 3D Stacked Memory-On-Logic", Proceedings of the IEEE Design, Automation, and Test in Europe Conference and Exhibition, pp. 1-6, 2014.

[43] G. Chandra, P. Kapur, and K. C. Saraswat, "A Methodology for the Interconnect Performance Evaluation of 2D and 3D Processors with Memory", Proceedings of the IEEE International Interconnect Technology Conference, pp. 164-166, 2002.

[44] W. Steinhoegl, G. Schindler, G. Steinlesberger, M. Traving, and M. Engelhardt, "Comprehensive study of the resistivity of copper wires with lateral dimensions of 100nm and smaller", IEEE Journal of Applied Physics, vol. 97, 023706-0237067, 2005.

[45] A. Ceyhan and A. Naeemi, "Cu interconnect limitations and opportunities for SWNT interconnects at the end of the roadmap," IEEE Transactions on Electron Devices, vol. 60, no. 1, pp. 374-382, Jan. 2013.

[46] I. Savidis and E. G. Friedman, "Closed-Form Expressions of 3-D Via Resistance, Inductance, and Capacitance", IEEE Transactions on Electron Devices", Vol. 56, pp. 1873-1881, 2009.

[47] J. A. Mandelman, R. H. Dennard, G. B. Bronner, J. K. DeBrosse, R. Divakaruni, Y. Li, and C. J. Radens, "Challenges and Future Directions for the Scaling of Dynamic Random Access Memory (DRAM)", IBM Journal of Research and Development, Vol. 46, Issue 2.3, pp. 187-212, 2002.

[48] W. Mueller, et al., "Challenges for the DRAM Cell Scaling to 40nm", Proceedings of the IEEE International Electron Devices Meeting, 2005, pp. 335-339.

[49] P. Kogge, et al., "Exascale computing study: technology challenges in achieveing exascale systems", Editor & Study Lead, 2008.

[50] Y. Huai, F. Albert, et al., “Observation of Spin-Transfer Switching in Deep Submicron-Sized and Low-Resistance Magnetic Tunnel Junctions”, Appl. Phys. Lett. 84, 3118, 2004.

[51] G. D. Fuchs, et al., “Spin-Transfer Effects in Nanoscale Magnetic Tunnel Junctions”, Appl. Phys. Lett. 85, 1205, 2004.

[52] J. Hayakawa, et al., “Current-Driven Magnetization Switching in CoFeB/MgO/CoFeB Magnetic Tunnel Junctions”, Appl. Phys. 44, L1267, 2005.

[53] J. M. Lee, L. X. Ye, M. C. Weng, Y. C. Chen, Simon C. Li, J. P. Su, Te-Ho Wu, “Spin Transfer Magnetization Switching Read/Write Cycle Test in MgO-Based Magnetic Tunnel Junctions”, IEEE Transactions on Magnetics, Vol. 43, No. 7, July 2007.

[54] S. K. Gupta, et al., Proceedings of the Design, Automation, and Test in Europe Conference and Exhibition, pp. 1455-1458, 2012.

[55] S. H. Kang, Qualcomm Inc., Non-Volatile Memories Workshop, 2013.

[56] R. Beach, et al., Proceedings of the IEEE IEDM, pp. 1-4, 2008.

[57] S. Byungkyu, et al., Proceedings of the International SOC Design Conference 2014.

[58] W. Li, and C. M. Tan, Proceedings of the International Conference of Electron Devices and Solid-State Circuits, pp. 1-2, 2011.

[59] Zhitao Diao, et al., J. Phys.: Condens. Matter. 19, 165209, 2007.

[60] H. Park, et al., Proceedings of the IEEE International Symposium on Nanoscale Architectures, pp. 53-58, 2011.

[61] H. Nazarian, “Crossbar resistive memory: the futute technology for NAND Flash”, Crossbar Inc., 2013.

[62] T. –Y. Liu, et al., “A 130.7mm2 2-layer 32Gb ReRAM memory device in 24nm technology”, IEEE International Solid-State Circuits Conference, pp. 210-211, 2013, .

[63] C. Yeh, et al., “Compact one-transistor-N-RRAM array architecture for advanced CMOS technology”, IEEE Journal of Solid-State Circuits, 50(5): 1299-1309, May 2015.

[64] S. Sheu, “A 4Mb embedded SLC resistive-RAM macro with 7.2ns read-write random-access time and 160ns MLC-access capability”, International Solid-State Circuits Conference, pp. 200-202, 2011.

[65] T. Morikawa, et al., “A low power phase change memory using low thermal conductive doped-Ge2Sb2Te5 with nano-crystalline structure”, IEEE International Electron Devices Meeting, pp. 31.4.1-31.4.4, 2012.

[66] M. Gottwald, et al., “Material development for perpendicularly magnetized tunnel junctions”, Non-volatile Memories Workshop, University of California San Diego, 2013.

[68] L. Zhang, et al., “High-drive current (> 1MA/cm 2) and highly nonlinear (> 10 3) TiN/amorphous-Silicon/TiN scalable bidirectional selector with excellent reliability and its variability impact on the 1S1R array performance”, IEEE International Electron Devices Meeting, pp. 6.8.1-6.8.4, 2014.